La Potenza è nulla senza controllo*

Last update: 08/07/2019

Massimiliano Pastore[1], Francesca Lionetti[2], Antonio Calcagnì[1], Gianmarco Altoè[1]

[1]Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova.
[2]Department of Biological and Experimental Psychology, Queen Mary University of London.

Sommario ll tema della replicabilità riveste oggi un ruolo centrale nel dibattito sull'attendibilità dei risultati delle ricerche in psicologia. Diverse sono ad oggi le possibili soluzioni proposte: la maggior cura nella definizione del disegno sperimentale alla luce delle ipotesi di ricerca, l'utilizzo di misure caratterizzate da ottimali livelli di affidabilità e validità, la registrazione degli studi e la loro valutazione da parte di esperti prima che i dati vengano raccolti e analizzati, le pratiche di condivisione dei dati e dei relativi codici utilizzati per le analisi. Accanto a queste raccomandazioni che stanno già portando a notevoli benefici, il dibattito sulla replicabilità ha riguardato naturalmente anche aspetti prettamente legati all'utilizzo della statistica. I temi più affrontati di recente sono la pianificazione della dimensione campionaria o power analysis e l'approccio bayesiano. In questo contributo, attraverso un semplice esempio, intendiamo sottolineare quanto, accanto ai temi già citati, la diffusione di una maggiore sensibilità statistica e una conseguente miglior scelta dei metodi di analisi giochino un ruolo cruciale, ma spesso sottovalutato, nella produzione di risultati maggiormente attendibili. In conclusione, crediamo che un aumento della formazione statistica nelle nuove generazioni di psicologi possa contribuire in modo determinante a migliorare la qualità della ricerca.

Parole chiave: replicabilità, affidabilità, validità, power analysis, significatività.

Power is nothing without control

Summary Replicability is currently in the spotlight of the debate around the reliability of research findings in psychology. Well-designed experiments, reliable and valid assessment measures, pre-registration of studies, open access to datasets and to statistic scripts used for running analyses are among the solutions proposed for overcoming the replicability crisis. Close to these methodological recommendations, power-analysis and the use of a Bayesian approach have been also advocated as candidate solutions to the replicability crisis in the psychology field. In this brief report, with a practical exemplification, we illustrate how a more extended knowledge of available statistical methods and approaches, and a more thoughtful choice of appropriate data analysis techniques, depending on the type of data, can contribute to more reliable research findings. To conclude, we propose that working for improving psychology researchers' statistical knowledge represents a pivotal way for promoting more reliable and replicable research findings, thus reducing the replicability crisis.

La corrispondenza va inviata a
Massimiliano Pastore
c/o Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Via Venezia, 8 I-35131 Padova (PD), Italy.
Email: massimiliano.pastore@unipd.it

* Please cite as: Pastore, M., Lionetti, F., Calcagnì, A., Altoè, G. (2019). *Giornale Italiano di Psicologia*, 46, 359–378.

1 INTRODUZIONE

Il tema della replicabilità e riproducibilità è entrato definitivamente nel dibattito sulla qualità della ricerca, compresa la psicologia; il numero 4 (vol. 13) di Luglio 2018 della rivista *Perspectives on Psychological Science* è quasi interamente dedicato a questo tema (Brainerd & Reyna, 2018; Fiedler, 2018; Frankenhuis & Nettle, 2018; Grand, Rogelberg, Banks, Landis, & Tonidandel, 2018; Kaufman & Glăveanu, 2018; Vazire, 2018; Wagenmakers, Dutilh, & Sarafoglou, 2018; Wai & Halpern, 2018). In ambito italiano, una recente rassegna del Giornale Italiano di Psicologia ha coinvolto vari autori che hanno discusso in particolare dei problemi connessi alla ricerca empirica in psicologia (Agnoli & Carollo, 2018; Perugini, 2018) mettendo in evidenza come un approccio *Open Science* e la condivisione dei dati possano fornire un utile supporto per migliorarne la qualità (Crepaldi, 2018; Grassi, 2018; Zogmaister, 2018).

La replicabilità implica che uno o più effetti individuati in una ricerca (sperimentale o quasi sperimentale) si possano riottenere, più o meno alle stesse condizioni, in altri campioni di dati (Whitaker, 2017). Si tratta di un indicatore importante, per non dire cruciale, della qualità della ricerca scientifica, forse oggi ancora di più, anche in relazione al significativo aumento di studi empirici ed articoli pubblicati in riviste scientifiche di alto impatto, esito di una crescente pressione a pubblicare output di ricerca "statisticamente significativi" a fini di avanzamenti di carriera e di vincita di finanziamenti per la ricerca. Ad un aumento della produzione scientifica, frutto del tanto discusso "publish or perish", non sembra essere corrisposto un aumento dell'affidabilitá dei prodotti scientifici (Smaldino & McElreath, 2016; Grimes, Bauch, & Ioannidis, 2018). Essere in grado di replicare un risultato scientifico è fondamentale per poter parlare di scientificità.

Il dibattito sulla qualità della ricerca è attivo da molto tempo (Ioannidis, 2005), ma recentemente ha coinvolto una quota sempre più rilevante di studiosi vicini alla ricerca psicologica ed alle scienze sociali più in generale (Bakker & Wicherts, 2011; Camerer et al., 2018; Francis, 2012; Munafò et al., 2017; Open Science Collaboration, 2015). Un punto fondamentale del dibattito, recentemente avanzato da più autori, è che la replicabilità non debba intendersi come semplice riproduzione di un risultato statisticamente significativo; la significatività statistica viene spesso male interpretata (si veda ad es: Cohen, 1994; McShane, Gal, Gelman, Robert, & Tackett, 2018; Wasserstein & Lazar, 2016; Ziliak & McCloskey, 2008) e dipende da molti elementi che entrano in gioco senza alcun legame con gli effetti studiati (Wagenmakers, 2007). Dunque, la significatività non può costituire il solo (o quantomento prevalente) criterio di valutazione della bontà del disegno di ricerca e del lavoro svolto, ossia un chiaro criterio di replicabilità.

Di fatto, uno dei problemi più rilevanti è la mancanza di potenza adeguata (*lack of power*; si veda ad es. Button et al., 2013; Cohen, 1962; Fraley & Vazire, 2014; Ioannidis, 2005; Lucas & Donnellan, 2013) che ha caratterizzato da molto tempo la ricerca psicologica. La stima a priori della potenza (Brysbaert & Stevens, 2018; Perugini, Gallucci, & Costantini, 2018) è stata proposta come una delle soluzioni elettive per far fronte alla questione replicabilità. Tuttavia, come vedremo in questo contributo, anche l'utilizzo della potenza come unico elemento di valutazione della possibilità di replicare i risultati è potenzialmente soggetto alla stessa fallacia di un approccio basato unicamente sul considerare il p-value. Come illustreremo in questo contributo, altre modalità di esplorazione dei dati possono essere maggiormente informative e garantire una accurata verifica della replicabilità.

Il concetto di potenza è molto importante sotto l'aspetto teorico quando si pianificano delle ricerche o esperimenti. Secondo il classico approccio Neyman-Pearson (Garthwaite, Jolliffe, & Jones, 2009) è la probabilità di ottenere un risultato significativo, ovvero rigettare l'ipotesi nulla, quando la stessa è falsa. Lo schema

previsto secondo tale approccio consiste nel fissare a priori il valore di α (probabilità di commettere un errore di I tipo) e quindi minimizzare β (probabilità di commettere un errore di II tipo) oppure, che è lo stesso, massimizzare la potenza $(1 - \beta)$. Gigerenzer e Marewski (2015), rifacendosi a Neyman (1957), evidenziano però che: *The usefulness of this procedure is limited among others to situations where there is a disjunction of hypotheses (e.g., either μ_1 or μ_2 is true), where there is repeated sampling, and where you can make meaningful cost-benefit trade-offs for choosing α and β.* In pratica, da questa frase possiamo rilevare facilmente alcune contraddizioni in cui ci si trova oggi seguendo lo schema ibrido, derivato dall'impropria fusione del metodo di Fisher con quello di Neyman e Pearson (Bachmann, Luccio, & Salvadori, 2005; Gigerenzer & Marewski, 2015; Pastore, 2009), che ha preso il nome di *Null Hypothesis Significance Testing* (NHST; Cohen, 1994); infatti in tale schema procedurale, la definizione dell'ipotesi alternativa (o ipotesi di ricerca) non è mai precisa e puntuale. Pertanto risulta piuttosto complicato il calcolo della potenza quando l'ipotesi H_1 è definita su un insieme di numeri reali. Ad esempio, se abbiamo $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, i valori possibili di θ sotto l'ipotesi alternativa sono infiniti e di conseguenza il calcolo di $1 - \beta$ si traduce nell'integrazione di infinite porzioni di densità di probabilità. Il calcolo diventa ancora più complicato nel caso di ipotesi multivariate o testing multiplo, dal momento che la potenza non è più univocamente definita (si veda ad es. Ramsey, 1979; Westfall & Young, 1993), o ancora, quando si fa riferimento a parametri che derivano da modelli complessi (modelli fattoriali o multilivello; si veda ad es. Chin, 1998; MacCallum, Browne, & Sugawara, 1996; Maas & Hox, 2005; Scherbaum & Ferreter, 2009). Nella prassi, la stima a priori della potenza viene utilizzata per stabilire una numerosità campionaria appropriata per il tipo ricerca. Nonostante questo abbia l'indubbio vantaggio di proteggere dal pericolo legato all'utilizzo di campioni troppo piccoli (falsi effetti significativi quindi difficilmente replicabili), il fatto che il termine potenza venga interpretato come se l'obiettivo di uno studio fosse solo quello di ottenere la significatività statistica (Gelman, 2017) può generare importanti confusioni.

Nell'esempio che segue, vogliamo dimostrare che, anche seguendo le migliori premesse ed avendo una potenza a priori adeguata, il risultato può essere distorto se non ci preoccupiamo anche di valutare quali siano le previsioni associate al modello statistico utilizzato.

Tutte le analisi statistiche sono state effettuate in ambiente R (R Core Team, 2018), servendoci dei pacchetti: `ez` (Lawrence, 2016), `lme4` (Bates, Mächler, Bolker, & Walker, 2015), `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2017) e `BayesFactor` (Morey & Rouder, 2018); per la grafica è stato utilizzato il pacchetto `ggplot2` (Wickham, 2009). Dati e codici R delle analisi statistiche presentate in seguito sono disponibili al link: `http://147.162.146.188/~pastore/data/nopower.R`

2 ESEMPIO EMPIRICO

Immaginiamo un esperimento in cui la variabile dipendente, che chiameremo Y, sia la proporzione di tempo dedicato alla fissazione di uno stimolo, oppure la proporzione di risposte corrette in una serie di prove. Questo tipo di misura è molto comune negli esperimenti in psicologia, ad esempio negli studi con i neonati o con animali, o nei compiti di riconoscimento. L'esperimento si compone di due fasi (f), una di training o abituazione, ed una di test vero e proprio, ciascuna composta da 10 prove (p). Nella fase di test ci sono due diverse condizioni sperimentali, A e B, che sono la parte cruciale dell'esperimento in quanto ci si attende che tra queste emergano delle differenze.

L'esperimento prevede di utilizzare un campione di soggetti da dividere casualmente in due gruppi, ciascuno dei quali sarà assegnato ad una sola delle due condizioni sperimentali (A oppure B). Dopo le due fasi sperimentali, per ciascun soggetto ci saranno pertanto 10 rilevazioni nella condizione di training ed altrettante

			training		test	
soggetto	condizione		media	dev.st	media	dev.st
1	A	A	0.69	0.09	0.77	0.04
2	B	A	0.69	0.11	0.76	0.03
3	C	A	0.61	0.20	0.62	0.10
4	D	A	0.50	0.07	0.78	0.06
5	E	A	0.61	0.25	0.76	0.06
6	F	A	0.52	0.16	0.69	0.05
7	G	A	0.61	0.02	0.69	0.08
8	H	A	0.72	0.28	0.59	0.10
9	I	B	0.68	0.24	0.49	0.45
10	J	B	0.71	0.22	0.81	0.14
11	K	B	0.37	0.29	0.74	0.32
12	L	B	0.53	0.37	0.88	0.07
13	M	B	0.61	0.28	0.90	0.13
14	N	B	0.63	0.22	0.86	0.11
15	O	B	0.67	0.32	0.97	0.06
16	P	B	0.58	0.21	0.62	0.11

Tabella 1. Medie (con dev. st.) della variabile Y per ciascuno dei 16 soggetti nelle due fasi sperimentali (dati simulati).

nella fase di test, 20 in tutto. L'ipotesi a priori è che si dovranno osservare, su Y, delle differenze tra i gruppi nella fase di test e non nella fase di training, quindi che ci sia un'interazione tra il fattore condizione ed il fattore fase.

2.1 Analisi di potenza

Prima di procedere con il reclutamento dei soggetti, facciamo un'analisi di potenza a priori al fine di stabilire il numero di soggetti da reclutare per questo studio. In particolare, dato che siamo interessati all'interazione, ci focalizziamo su questo effetto. Da conoscenze pregresse o dalle nostre ipotesi basate sulla letteratura del settore[2], ci attendiamo un effect size per l'interazione di circa 0.20 (espresso con un valore di η^2 generalizzato; Bakeman, 2005). Adottando una procedura di simulazione Monte Carlo[3], facendo variare la numerosità campionaria tra 6 e 100 otteniamo che la potenza stimata con 16 soggetti è circa del 73%. Quindi, facendo una debita considerazione di costi e benefici, supponiamo di considerare accettabile questo livello di potenza, e procediamo con il reclutamento di 16 partecipanti che invitiamo a prendere parte all'esperimento.

2.2 ANOVA in senso tradizionale

Al termine dell'esperimento, la prassi comunemente seguita consiste nel calcolare le medie su Y di ciascun soggetto nelle due fasi e nelle due condizioni sperimentali e quindi analizzare con un modello ANOVA a misure ripetute l'effetto dell'interazione tra fase e condizione. In tabella 1 sono riportate medie e deviazioni standard di Y per i 16 soggetti dei due gruppi sperimentali nelle due fasi (i dati sono simulati). In figura 1, pannello $[A]$, sono rappresentate graficamente le medie complessive di Y nei due gruppi sperimentali e nelle due fasi.

[2] Va ricordato che in generale la letteratura nelle scienze psicologiche e sociali tende a sovrastimare gli effetti (si veda ad es. Open Science Collaboration, 2015) e pertanto bisogna sempre valutarla con attenzione.

[3] I dettagli della procedura sono disponibili su richiesta al primo autore, in questa sede li omettiamo per non dilungarci troppo.

Il risultato dell'analisi della varianza è riportato in tabella 2. Nella porzione sinistra ci sono le classiche informazioni che vengono riportate nelle tabelle ANOVA ovvero la statistica F con relativi gradi di libertà (df_n e df_d) e *p-value*. Nella parte destra della stessa tabella abbiamo aggiunto le informazioni necessarie al calcolo del Bayes Factor (BF). Seguendo le indicazioni di Masson (2011) abbiamo prima calcolato il BIC (*Bayesian Information Criterion*; Schwarz, 1978) utilizzando la seguente formula:

$$\text{BIC} = n \log(1 - R^2) + k \log(n)$$

in cui $1-R^2$ è la variabilità non spiegata dal modello data dal rapporto tra la devianza residua ($SS\text{e}$) e quella totale (SS_{tot}), $\frac{SS\text{e}}{SS_{tot}}$, $n = s(c-1)$ il numero di osservazioni indipendenti, in cui s è il numero di soggetti e c il numero di condizioni in cui i soggetti sono ripetuti, e k il numero di parametri del modello. Poi abbiamo utilizzato le differenze tra il BF del modello nullo e gli altri per calcolare i Bayes Factor associati agli effetti del modello con la formula $exp(\Delta_{\text{BIC}}/2)$ (Raftery, 1995).

	df_n	df_d	F	p	SSe	SStot	BIC	Δ_{BIC}	BF
modello nullo							2.77	0.00	1.00
condizione	1	14	0.53	0.478	0.17	0.17	2.18	0.60	1.35
fase	1	14	11.90	0.004	0.18	0.33	-7.07	9.84	137.24
condizione:fase	1	14	1.51	0.240	0.18	0.20	1.14	1.64	2.27

Tabella 2. Tabella ANOVA relativa all'esperimento ($n = 16$): df = gradi di libertà (n = numeratore, d= denominatore), SSe = devianza residua, SStot = devianza totale, BIC = *Bayesian Information Criterion*, BF = *Bayes Factor* rispetto al modello nullo.

Dalla lettura della tabella osserviamo che l'unico effetto significativo (con $p < .05$) è quello della fase, ovvero le differenze associate ai valori osservati nella fase di training e nella fase di test. Il fattore fase presenta un BF di circa 137, che indica un effetto molto plausibile. Il fattore condizione non risulta significativo ed il BF quasi uguale ad 1 supporta l'ipotesi di non differenza tra i gruppi nelle due condizioni sperimentali. Infine, l'interazione non è statisticamente significativa e presenta un BF di circa 2 a supporto di un effetto plausibile, ma solo di poco. Pertanto, sulla base di questo risultato, l'ipotesi di interazione non sembrerebbe

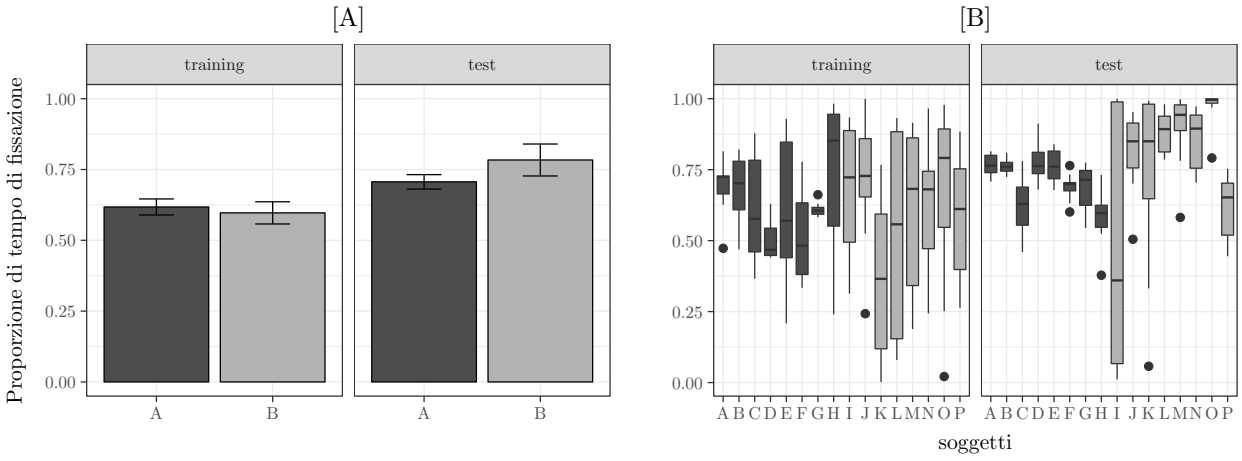

Figura 1. Risultati dell'esperimento. [A] rappresentazione a barre delle medie di Y (con errore standard) nei due gruppi e nelle due fasi; [B] distribuzione dei valori di Y per ciascun soggetto separatamente.

trovare conferma. Qualora si trattasse di uno studio di replica, la conclusione sarebbe che l'esperimento non regge alla replica, anche avendo coinvolto un numero adeguato di soggetti come testimoniato dall'analisi di potenza a priori.

2.3 Mixed model

Il maggiore limite del modello appena utilizzato è che non tiene in debita considerazione la variabilità individuale. Se osserviamo il pannello [B] della figura 1, in cui sono rappresentate con dei boxplot le distribuzioni dei valori osservati di Y per ogni soggetto nelle 10 prove, vediamo come le apparenti similitudini medie della tabella 1 nascondano forti differenze interindividuali e suggeriscano la necessità di individuare un modello di analisi più appropriato, che ne tenga conto.

Per fare questo utilizziamo quindi un *mixed effects model* (MEM; Pinheiro & Bates, 2000) in cui i valori osservati dai soggetti in ciascuna prova vengono considerati separatamente (senza calcolare le medie per soggetto e per fase), quindi per un totale di 16 (soggetti) $\times$ 10 (prove) $\times$ 2 (fase) = 320 valori di Y rilevati. I fattori del modello sono gli stessi di prima, condizione sperimentale, fase e la loro interazione, con l'aggiunta dei soggetti come fattore random.

	df_n	df_d	F	p	BF
condizione	1	14	0.53	0.48	0.22
fase	1	302	35.49	0.00	390806.29
condizione:fase	1	302	4.50	0.03	1.18

Tabella 3. Tabella ANOVA relativa all'esperimento basata su un *mixed model* in cui i soggetti ($n = 16$) sono trattati come effetti random: df = gradi di libertà (n = numeratore, d= denominatore), BF = *Bayes Factor* rispetto al modello nullo.

In tabella 3 sono riassunti i risultati di questa nuova analisi. Notiamo che in questo caso l'effetto di interazione è diventato significativo ($p = 0.03$). Se ci fermassimo a livello di significatività, le conclusioni che trarremmo in questo caso sarebbero diverse: l'ipotesi verrebbe supportata, e in caso si trattasse di uno studio di replica, la conclusione sembrerebbe positiva. Tuttavia, da una analisi piu approfondita, notiamo che il Bayes Factor associato all'effetto interazione rimane comunque piuttosto basso (BF = 1.18). La sostanza delle conclusioni non è molto diversa da quelle basate sulla tabella 2 e sull'analisi della varianza: abbiamo un effetto di interazione poco supportato ed una differenza molto evidente tra i valori nelle due fasi, mentre la condizione sperimentale non sembra incidere molto.

Sicuramente l'avere tenuto in considerazione le differenze individuali ha fornito dei risultati più completi. Ricordiamo anche che, tra i parametri del modello appena considerato, vi è anche la deviazione standard dei soggetti, stimata in circa 0.06. Questo valore è una misura della variabilità nella fase di training e quindi, il fatto che sia piuttosto basso, suggerisce che in questa fase non vi siano particolari differenze tra i soggetti, come dovrebbe essere.

2.4 Cosa prevede il modello?

In entrambi i tipi di analisi utilizzata, eseguiti a seguito del calcolo della potenza, sono emerse potenziali fonti di distorsione che in un approccio dicotomico potrebbero potenzialmente portare a conclusioni non affidabili. Se fossimo in una condizione di replica, sarebbe difficile trarre conclusioni certe ed il rischio di errore sarebbe alto. Focalizziamoci allora su un aspetto che viene spesso omesso o non considerato, ovvero su cosa prevede

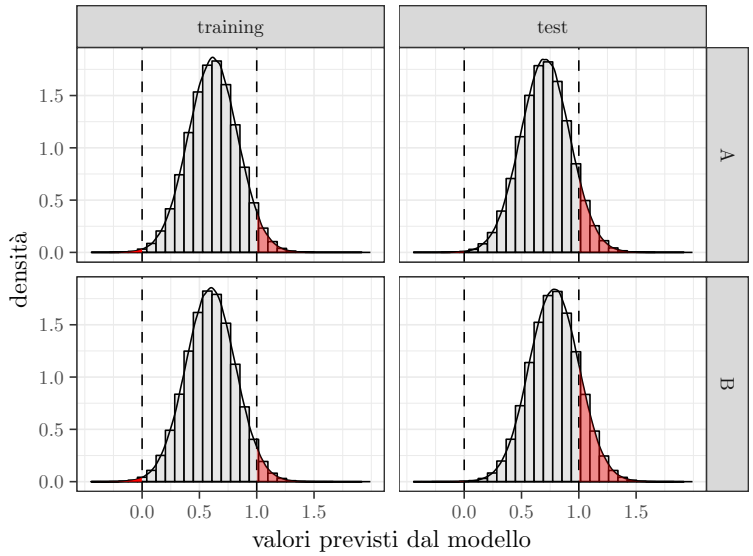

Figura 2. Distribuzione dei valori previsti dal modello in 10000 campioni simulati di dati sulla base dei parametri stimati nelle due condizioni sperimentali (A, sopra, e B, sotto) e nelle due fasi (training, a sinistra, e test, a destra). Le linee tratteggiate sono poste a 0 e 1, soglie all'interno delle quali si trovano i valori effettivamente osservabili e le aree evidenziate sono relative ai valori non plausibili empiricamente.

il modello. I modelli statistici infatti, oltre a cercare di interpretare i dati che osserviamo, servono anche per fare delle previsioni su cosa attenderci qualora dovessimo osservare nuovi dati. Questo aspetto assume particolare importanza se siamo interessati alla replicabilità.

Per avere un'idea di cosa prevede il modello, simuliamo 10000 campioni di dati sulla base dei parametri e degli effetti stimati sul campione empirico. In figura 2 sono rappresentate le distribuzioni di valori ottenuti di Y nelle due condizioni sperimentali (A, sopra, e B, sotto) e nelle due fasi (training, a sinistra, e test, a destra). La figura mette subito in evidenza un problema di questo modello: dal momento che la variabile dipendente è una proporzione, essa deve per forza assumere valori compresi nell'intervallo $[0 - 1]$. Sfortunatamente il modello prevede una non trascurabile percentuale di casi che escono da questo intervallo e precisamente il 4% nella condizione A, fase di training, il 3% nella condizione B, fase di training, il 9% nella condizione A, fase di test, ed infine il 16% nella condizione B, fase di test; il modello prevede complessivamente nelle fasi di test una percentuale di valori di circa 12% che sarebbe impossibile osservare. La ragione di questo è che per testare gli effetti, sia nell'esempio ANOVA sia nell'esempio MEM, abbiamo utilizzato un modello lineare che si aspetta: 1) una variabile dipendente quantitativa che può assumere qualunque valore reale tra meno e più infinito e 2) dei residui normalmente distribuiti. In sintesi anche ricorrendo ad un MEM, e nonostante questo abbia permesso di tenere in debita considerazione la variabilitá individuale, non é stata considerata la vera natura dei dati.

2.5 Modello binomiale e confronto tra modelli

Quando la variabile dipendente è costretta naturalmente nell'intervallo $[0 - 1]$, il modello da utilizzare è quello logistico. In pratica si tratta di adottare, al posto del modello lineare semplice, un modello lineare generalizzato in cui viene inclusa una funzione legame di tipo logistico che associa i valori osservati delle proporzioni (o della variabile Y) ai predittori del modello, ma in modo che la variabile dipendente non venga

	χ^2	df	p
condizione	1.06	1	0.30
fase	67.72	1	0.00
condizione:fase	11.65	1	0.00

Tabella 4. Tabella degli effetti dell'esperimento basata su un *generalized mixed model* in cui i soggetti ($n = 16$) sono trattati come effetti random: df = gradi di libertà.

snaturata. Il modello logistico consente non solo di stimare meglio i parametri, e di conseguenza gli effetti, ma anche di non avere valori previsti fuori dal range di valori osservabili empiricamente.

In tabella 4 è riportato l'esito dell'analisi sugli stessi dati utilizzando un modello lineare generalizzato; gli effetti sono riassunti nella tradizionale tabella di Analisi della Devianza. Il problema di questa tabella è che fornisce informazioni circa la significatività degli effetti ma non sulla loro grandezza (il valore di p non permette di quantificare l'evidenza statistica; si veda ad es. Berger & Sellke, 1987; Wagenmakers, 2007; Ziliak & McCloskey, 2008). Per poter stimare opportunamente le grandezze degli effetti dobbiamo cambiare leggermente la modalità di analisi. In pratica, al posto della tradizionale tabella ANOVA (o Analisi della Devianza in questo caso specifico), applicheremo una strategia di confronto tra modelli (Burnham & Anderson, 2003; McElreath, 2016; Wagenmakers & Farrell, 2004) basata sul rapporto di verosimiglianza (*Likelihood Ratio Test*; Mood & Graybill, 1963).

L'idea alla base di questa strategia è piuttosto semplice. Al posto di una analisi unica in cui si inseriscono direttamente tutti gli effetti che ci interessano (condizione, fase e interazione), costruiamo dei modelli in cui introduciamo in sequenza un effetto alla volta. Ad esempio, si può iniziare con il modello nullo in cui la variabile dipendente è spiegata solo dall'effetto random soggetti. Quindi introduciamo nel modello la variabile condizione, poi aggiungiamo la variabile fase, così da ottenere il modello additivo, ed infine aggiungiamo l'interazione condizione × fase.

Nella parte sinistra della tabella 5 vediamo il risultato di questa procedura. Ciascuna riga della tabella è riferita ad uno dei quattro modelli considerati. Le prime due colonne (df_m e dev.) riportano rispettivamente gradi di libertà e devianza del modello. La devianza misura la quota residua (o non spiegata) di variabilità del modello e pertanto sarà massima nel modello nullo e tenderà a diminuire mano a mano che si aggiungono nuovi effetti nel modello. La differenza tra le devianze di due modelli nidificati ha una distribuzione campionaria approssimabile con un χ^2 con gradi di libertà pari alla differenza dei gradi di libertà tra i due modelli. Queste due informazioni sono riportate nella terza e quarta colonna della tabella, nella quinta è riportato il relativo *p-value*. Quest'ultimo ci permette di valutare se la differenza tra due modelli risulti statisticamente significativa o, in altri termini, se l'aggiunta di un determinato effetto nel modello comporti un miglioramento significativo nell'adattamento ai dati.

In pratica, nella parte sinistra della tabella 5 vediamo che l'effetto condizione non è significativo ($p = .3$) mentre gli altri effetti sono statisticamente significativi (con $p < .01$). Questo risultato è perfettamente

	df_m	dev.	χ^2	df_{χ^2}	p	BIC	Δ_{BIC}	logBF	weight
M0: modello nullo	2	1552.8				1564.34	0.00	0.00	0.00
M1: condizione	3	1551.7	1.09	1	0.30	1569.02	-4.68	-2.34	0.00
M2: condizione + fase	4	1481.3	70.40	1	0.00	1504.39	59.95	29.97	0.05
M3: condizione × fase	5	1469.6	11.70	1	0.00	1498.46	65.88	32.94	0.95

Tabella 5. Tabella del confronto tra modelli lineari generalizzati con effetto random (i soggetti) e funzione legame logit: df = gradi di libertà (m = modello), dev. = devianza del modello, χ^2 = chi-quadrato, BIC = *Bayesian Information Criterion*, logBF = logaritmo del *Bayes Factor* rispetto al modello nullo, weight = Akaike weight.

coerente con quello già visto in tabella 4. Ma la strategia di confronto tra modelli ci consente di avere delle informazioni in più, come vediamo nella parte destra della tabella. Oltre ai già citati BIC, Δ_{BIC} e BF (qui espresso in logaritmi per evitare numeri enormi), nell'ultima colonna (weight) abbiamo una distribuzione di probabilità, con un valore per ciascun modello. I valori di questa colonna si ottengono semplicemente normalizzando i Δ_{BIC}[4] e si possono interpretare come *la probabilità che il modello sia il più verosimile tra i modelli scelti sulla base dei dati osservati* (per maggiori dettagli si veda: Burnham, Anderson, & Huyvaert, 2011; McElreath, 2016; Wagenmakers & Farrell, 2004). La conclusione è che, utilizzando un appropriato metodo per trattare i dati dell'esperimento, il modello con interazione risulta essere il più evidente con una probabilità di circa 19 volte superiore a quella del modello senza interazione. Aver quantificato in modo probabilistico l'effetto interazione, ha permesso di avere maggiori informazioni dai dati e di ridurre il rischio di distorsione, garantendo maggior affidabilità nelle conclusioni che dall'esperimento possiamo trarre. Permetterà inoltre, in fase di replica da parte di un laboratorio/gruppo di ricerca diverso e indipendente, di quantificare in che misura i nuovi risultati si discostino dallo studio originale con un indicatore di facile ed intuitiva interpretazione. Ad esempio: in questo esperimento, il modello interazione è 19 volte piú probabile del modello con gli effetti additivi. Nell'esperimento di replica, il modello di interazione potrebbe essere n volte piú/meno probabile del modello a effetti principali.

A questo punto possiamo concentrare la nostra attenzione sul modello migliore (M3) andando ad ispezionare graficamente le attese del modello. Come vediamo in figura 3, l'interazione appare chiaramente nelle proporzioni attese per le due condizioni. Nel gruppo della condizione A si passa da una proporzione di 0.62 nella fase di training ad una di 0.71 nella fase di test, mentre nel gruppo B abbiamo 0.61 nella fase di training e 0.8 nella fase di test.

Poiché il modello è logistico, possiamo calcolare la grandezza degli effetti utilizzando gli Odds Ratio (OR) sfruttando la relazione: $OR = exp(\beta)$ in cui β è un parametro del modello. In particolare, dato il valore

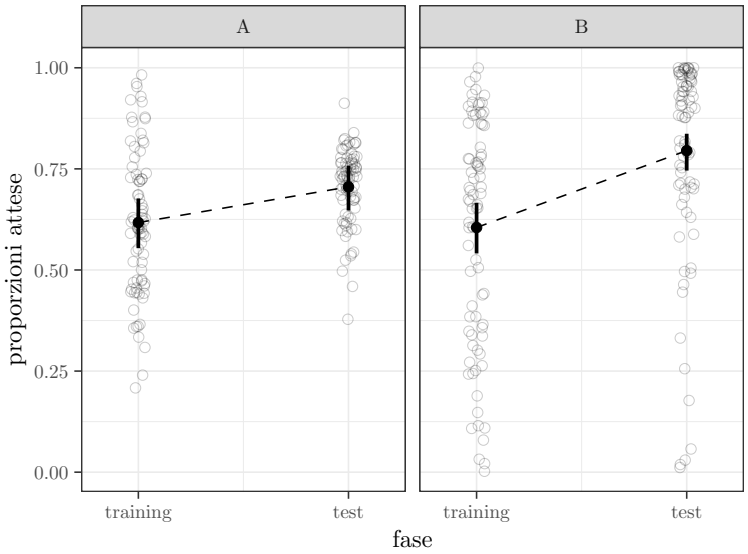

Figura 3. Medie della variabile Y attese dal modello M3 rispetto ai due gruppi e nelle due fasi. I punti grigi sono i valori osservati ($n = 320$), i segmenti sono gli intervalli di confidenza dei valori attesi.

[4] L'operazione di calcolo dei weight è la seguente $w_i = exp(\Delta_i)/\sum_i exp(\Delta_i)$ in cui Δ_i sono le differenze tra i criteri di informazione dei modelli considerati. In questo caso il criterio di informazione usato è BIC, ma si può procedere allo stesso modo con AIC ed altri criteri di informazione.

stimato del parametro di interazione $\beta = 0.53$, abbiamo che $exp(0.53) = 1.71$. Con dati di questo tipo, l'OR è particolarmente indicato per descrivere gli effetti. In questo caso specifico il valore associato all'interazione (1.71, con IC al 95% $[1.26, 2.32]$) rappresenta il rapporto tra due OR (Jaccard, 2001), ossia quello relativo al confronto post-pre nel primo gruppo vs quello relativo al confronto post-pre nel secondo gruppo. In altri termini, e nello specifico di questo esempio: l'OR del confronto tra la fase di training e quella di test è 1.71 volte maggiore nella condizione B, rispetto alla condizione A. In figura 3 possiamo visualizzare graficamente questo effetto; nei due pannelli sono rappresentate le proporzioni attese nelle due condizioni (A e B) rispetto alle due fasi (training e test) con il relativo intervallo di confidenza al 95%. I punti grigi sono le effettive osservazioni: 16 (soggetti) $\times$ 10 (prove) $\times$ 2 (fase) = 320.

Il valore ottenuto (1.71) ed il relativo intervallo di confidenza ($[1.26, 2.32]$) possono essere considerati una stima migliore dell'effetto di interazione rispetto a quella ricavata da un semplice modello lineare ed essere pertanto utilizzati per analisi di potenza a priori in nuovi studi o esperimenti.

3 NOTE CONCLUSIVE

Negli ultimi anni all'interno delle discipline psicologiche vi è stato un ampio dibattito metodologico sul tema della replicabilità dei risultati e della loro affidabilità. La pressione a pubblicare, e quasi solo risultati statisticamente significativi, sembra sia andata di pari passo con quella che da piú parti é stata evocata come una vera e propria "crisi". Diverse sono state le soluzioni proposte a questo annoso problema, da un punto di vista statistico e metodologico. Tra queste, l'analisi della potenza ha un posto di rilievo. Molte sono oramai le riviste che richiedono, in fase di sottomissione dei lavori, un'opportuna giustificazione della definizione campionaria mediante un'analisi della potenza. Tuttavia ciò non ha evitato ai ricercatori un utilizzo inopportuno di questo metodo che è stato invece ampiamente impiegato per ottenere risultati significativi (più che informazioni a supporto della definizione campionaria; si veda ad. es.: Benjamin et al., 2018; Lakens et al., 2018). Nonostante questo strumento rivesta un ruolo indubbiamente importante dal punto di vista metodologico e abbia avuto il merito di porre l'attenzione sull'esigenza di una adeguata numerosità campionaria per trarre informazioni affidabili dai dati, occorre sottolineare come lo studio della potenza non garantisca da sé una tutela da possibili distorsioni nell'interpretazione dei risultati e rischi inoltre di calcare la mano, ancora una volta, sul concetto di significativitá statistica come unico criterio per valutare la bontá di un output di ricerca.

Una soluzione proposta in letteratura, e che in questo lavoro abbiamo descritto con alcune esemplificazioni pratiche, è quella di passare invece a metodi maggiormente focalizzati sugli effetti più che sui semplici *p-value* (Amrhein & Greenland, 2018; Gelman & Carlin, 2014; Ioannidis, 2018; McShane et al., 2018; Trafimow et al., 2017). Diverse sono infatti le voci critiche che suggeriscono di andare oltre un approccio basato unicamente sul *p*-value affinché si possa garantire una miglior comprensione dei dati unitamente ad una maggiore solidità nelle conclusioni che si possono trarre (Coyne, 2016; Gelman, 2016; Robinson, in stampa). Questo rappresenta un punto di svolta che solleva anche la questione su una più adeguata formazione statistica di chi fa ricerca in psicologia e più in generale nelle scienze sociali (Leek et al., 2017; Sharpe, 2013). A nostro parere una più approfondita cultura statistica contribuirebbe anche a formare una maggiore consapevolezza sull'uso (e abuso) delle metodologie e tecniche dell'analisi dei dati, in una direzione che contribuisca anche a sanare quella crisi di replicabilità da molti evocata. Un altro aspetto che riteniamo rilevante in questo dibattito è anche quello di una adeguata consapevolezza circa gli errori decisionali che un uso scorretto della statistica inferenziale può portare. Ci riferiamo non soltanto alla classica distinzione tra errore di primo tipo ed errore

di secondo tipo, altresì a quell'errore che si commette quando si sovrastima un effetto rilevante per la ricerca, un errore noto in statistica come *errore di tipo M* (Gelman & Carlin, 2014).

In un contesto complesso come la ricerca in psicologia, non esistono nuovi miti e soluzioni a priori. Anche nuovi metodi di analisi statistica, come quelli che si basano sulla statistica bayesiana, possono essere soggetti allo stesso tipo di distorsioni dei classici metodi di analisi basati sull'approccio NHST. Allo stesso modo l'analisi della potenza non garantisce conclusioni scevre da errore, laddove i dati non siano trattati rispettando la loro natura e le tecniche di analisi applicate in modo acritico e meccanicistico. Per questo proponiamo che un ruolo rilevante nella crisi di replicabilitá sia svolto da un utilizzo consapevole delle tecniche statistiche, sapendo che le diverse metodologie impiegate per valutare i propri risultati (NHST, analisi della potenza, valutazione della capacità previsionale del modello) forniscono informazioni ed evidenze *a supporto delle decisioni* del ricercatore, più che incontrovertibili oracoli circa l'esistenza o meno di determinati effetti, e come tali vanno valutate tenendo in debita considerazione il tipo di dati di volta in volta sotto esame; in sintesi, accogliere la variabilità e accettare l'incertezza (Gelman, 2015). Come sostengono Leek e colleghi, infatti, "data analysis is not purely computational and algorithmic - it is a *human behaviour*" (Leek et al., 2017, corsivo nostro).

Riferimenti bibliografici

Agnoli, F., & Carollo, G. (2018). Uso e (abuso) di prassi di ricerca problematiche in psicologia. *Giornale Italiano di Psicologia*.

Amrhein, V., & Greenland, S. (2018). Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, *2*, 4.

Bachmann, C., Luccio, R., & Salvadori, E. (2005). *La verifica della significatività dell'ipotesi nulla in psicologia.* Firenze University Press.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E., Berk, R., ... others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Berger, J. O., & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, *82*(397), pp. 112–122.

Brainerd, C., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science*, *13*(4), 428–432.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *Journal of Cognition*, *1*(1), 1–20.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

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.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... others (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 1–8.

Chin, W. W. (1998). Commentary: Issues and opinion on structural equation modeling. *MIS Quarterly*, *22*(1), vii–xvi.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *28*(4), 1–11.

Crepaldi, D. (2018). Open Science, Fair Science: Garantire la trasparenza della scienza attraverso l'organizzazione della pratica quotidiana in laboratorio. *Giornale Italiano di Psicologia*.

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*(4), 433–438.

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). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594.

Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science*, *13*(4), 439–447.

Garthwaite, P., Jolliffe, I., & Jones, B. (2009). *Statistical inference (Second Edition)*. Oxford University Press.

Gelman, A. (2015). Statistics and the crisis of scientific replication. *Significance*, *12*(3), 23–25.

Gelman, A. (2016). *Replication crisis crisis: Why I continue in my "pessimistic conclusions about reproducibility"*. Retrieved from `https://andrewgelman.com/2016/03/05/29195/` (Statistical Modeling, Causal Inference, and Social Science [Blog Post])

Gelman, A. (2017). *The "80% power" lie.* Retrieved from `https://andrewgelman.com/2017/12/04/80-power-lie/` (Statistical Modeling, Causal Inference, and Social Science [Blog Post])

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440.

Grand, J. A., Rogelberg, S. G., Banks, G. C., Landis, R. S., & Tonidandel, S. (2018). From outcome to process focus: Fostering a more robust psychological science through registered reports and results-blind reviewing. *Perspectives on Psychological Science*, *13*(4), 448–456.

Grassi, M. (2018). Crisi della riproducibilità e Open Science: alle porte di un cambio di paradigma. *Giornale Italiano di Psicologia*.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society open science*, *5*(1), 171511.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), 696–701.

Ioannidis, J. P. (2018). The proposal to lower p value thresholds to .005. *JAMA*, *319*(14), 1429-1430.

Jaccard, J. (2001). *Interaction effects in logistic regression.* SAGE, Thousand Oaks, CA.

Kaufman, J. C., & Glăveanu, V. P. (2018). The road to uncreative science is paved with good intentions: Ideas, implementations, and uneasy balances. *Perspectives on Psychological Science*, *13*(4), 457–465.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... others (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171.

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ez` (R package version 4.4-0)

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., & Goodman, S. N. (2017). Five ways to fix statistics. *Nature*, *551*(7682), 557–559.

Lucas, R. E., & Donnellan, M. B. (2013). Improving the replicability and reproducibility of research published in the Journal of Research in Personality. *Journal of Research in Personality*, *47*(4), 453–454.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130–149.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research*, *43*(3), 679–690.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC Press.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2018). Abandon statistical significance. *arXiv preprint arXiv:1709.07588v3*.

Mood, A., & Graybill, F. (1963). *Introduction to the Theory of Statistics* (2nd ed.). McGraw-Hill, New York.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-4.2)

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.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review*, *25*, 7–22.

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

Pastore, M. (2009). I limiti dell'approccio NHST e l'alternativa Bayesiana. *Giornale Italiano di Psicologia*, *36*(4), 925–938.

Perugini, M. (2018). Separare il segnale dal rumore: Alcuni suggerimenti. *Giornale Italiano di Psicologia*.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, *31*(1), 1–23.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus.* Springer.

R Core Team. (2018). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Ramsey, P. H. (1979). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *363*, 479–485.

Robinson, G. K. (in stampa). What Properties Might Statistical Inferences Reasonably be Expected to Have? -Crisis and Resolution in Statistical Inference. *The American Statistician*, *0*(0), 1–10.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347–367.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological methods*, *18*(4), 572.

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

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C., Beh, E. J., Bilgiç, Y., ... others (2017). Manipulating the alpha level cannot cure significance testing–comments on "Redefine statistical significance". *PeerJ Preprints*, *5*, e3411v1.

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

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The Creativity-Verification Cycle in Psychological Science: New Methods to Combat Old Idols. *Perspectives on Psychological Science*, *13*(4), 418–427.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, *11*(1), 192–196.

Wai, J., & Halpern, D. F. (2018). The Impact of Changing Norms on Creativity in Psychological Science. *Perspectives on Psychological Science*, *13*(4), 466–472.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing.* Wiley, NY.

Whitaker, K. (2017, September). Publishing a reproducible paper whitaker. In *Open science in practice summer school.* figshare. Retrieved from `https://doi.org/10.6084/m9.figshare.5440621.v2`

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance.* Ann Arbor, MI: University of Michigan Press.

Zogmaister, C. (2018). La condivisione dei dati deve diventare prassi comune per la ricerca psicologica. *Giornale Italiano di Psicologia*.