UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di laurea Magistrale in Psicologia dello sviluppo e dell'educazione

Tesi di laurea Magistrale

L'introduzione della preregistrazione nella ricerca psicologica: cause, benefici e limiti

Introduction to preregistration in psychological research: causes, benefits, and limits

Relatore
Prof.ssa Franca Agnoli

Laureando/a: **Marco Lezcano**
Matricola:**1232166**

Anno Accademico 2020/2021

Indice

Introduzione

Tra la fine della prima e l'inizio della seconda decade degli anni duemila una serie di eventi e pubblicazioni ha minato la stabilità e credibilità della ricerca scientifica nell'ambito della psicologia. In questa breve introduzione si descriveranno quali siano stati questi avvenimenti in modo da poter poi presentare gli aspetti più rilevanti della preregistrazione, tema centrale della presente tesi.

Prima di iniziare a descrivere gli eventi sopracitati è opportuno fare una precisazione relativa alle terminologie utilizzate per identificare questo periodo e che verranno poi riprese più avanti nel testo. Facendo riferimento all'articolo "*Psychology's Renaissance"* di Nelson, Simmons, e Simmonsohn (2017) viene identificata come *Replication Crisis* il periodo antecedente al 2010 dove ancora non si era a conoscenza delle problematiche relative alle metodologie di ricerca. Viene invece identificato come *Psychology's Renaissance* il periodo successivo, nel quale vi è la presa di coscienza da parte degli studiosi delle problematiche relative alle metodologie di ricerca e i criteri di pubblicazione. Così facendo inizieranno poi ad essere messe in atto una serie di accorgimenti volti a ripristinare la credibilità del campo di ricerca. Va anche menzionata la terminologia data da Vazire nel 2018, la quale utilizza il termine *Credibility Revolution* per identificare sia il periodo antecedente ai cambiamenti che quello successivo (Vazire, 2018). Nonostante la definizione, proposta da Vazire, sia più inclusiva in termini di etica e trasparenza, in questo elaborato verranno utilizzati i termini proposti da Nelson et al. (2017), poiché rendono più facile la distinzione tra prima e dopo l'attuazione delle riforme permettendo di fare meno confusione tra i due periodi.

I principali avvenimenti che dalla *Replication Crisis* orientano un percorso di analisi interna al modo in cui viene fatta ricerca in psicologia sono di seguito descritti.

Nel 2011 il ricercatore Diederik Stapel viene accusato di manipolazione e fabbricazione di dati da alcuni suoi dottorandi con cui collaborava alll'Università di Tilburg in Olanda. A seguito di indagini e verifiche di molti dei suoi articoli, quest'ultimi vengono ritirati da numerose riviste scientifiche, partendo da pubblicazioni risalenti al 2004. Infine, lo studioso venne allontanato

dall'università. Nello stesso periodo altri due casi di fabbricazione di dati e frode emersero, quello di Lawrence Sanna e Dirk Smeesters.

Nello stesso anno (2011) Daryl Bem pubblica l'articolo *Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect* sul *Journal of Personality and Social Psychology*. Con questo articolo Bem vuole dimostrare come eventi futuri non ancora avvenuti potessero avere delle influenze su eventi relativi ad un tempo antecedente. La conferma delle ipotesi sostenute e il fallimento di tutti i tentativi di replica (Wagenmakers, Wetzels, Borsboom, & Maas, 2011; Ritchie, Wiseman, & French, 2012; Galak, LeBoeuf, Nelson, & Simmons, 2012) della ricerca incominciarono a far riflettere le/i ricercatrici/ricercatori su come venisse condotta la ricerca all'interno del campo della psicologia sociale.

Un altro momento importante è la pubblicazione dell'articolo di John et al. (2012), nel quale i ricercatori cercarono di stimare nella maniera più accurata possibile l'uso delle *Questionable Research Practices* (QRP) all'interno dei diversi ambiti di ricerca della psicologia. Lo studio ha visto la presenza di 2155 partecipanti, i quali sono stati contattati tramite e-mail ed a cui è stato chiesto di fornire una misura di quanto avessero utilizzato ogni singola QRP elencata nel questionario, quanto fosse difendibile l'utilizzo di quella pratica, dare una stima di quanto venissero utilizzate da altri psicologi e quanti di questi avrebbero ammesso di metterla in atto. I partecipanti sono stati divisi in due condizioni. In una venivano proposti degli incentivi se si era stati sinceri nelle risposte (*bayesian-truth-serum condition* o BTS), mentre l'altra non aveva alcuna manipolazione. Dai risultati della ricerca emerge come il 94% dei partecipanti nella condizione BTS ammetta di aver utilizzato almeno una delle QRP elencate, rispetto ad un 91% nell'altra condizione. Inoltre, coloro che ammettevano di aver utilizzato queste pratiche tendevano poi a difendere e giustificare il loro comportamento. Per quel che riguarda la stima dell'utilizzo delle QRP da parte di altri ricercatori, questa era più alta rispetto a quella dell'ammissione per quasi tutte le singole pratiche. Inoltre, il 35% dei partecipanti riportava di aver avuto dei dubbi sull'integrità dei loro colleghi del proprio campo di ricerca in più

occasioni. Di seguito (Figura 1) è riportato un grafico presente all'interno dell'articolo. Il tasso di ammissione dell'uso, la stima della prevalenza e la stima della prevalenza derivata dalla stima del tasso di ammissione (tasso di ammissione / stima del tasso di ammissione) per ogni singola QRP indagata durante la ricerca nella condizione BTS sono di seguito rappresentati. I numeri sopra le barre delle QRP corrispondono alla media geometrica delle tre percentuali.

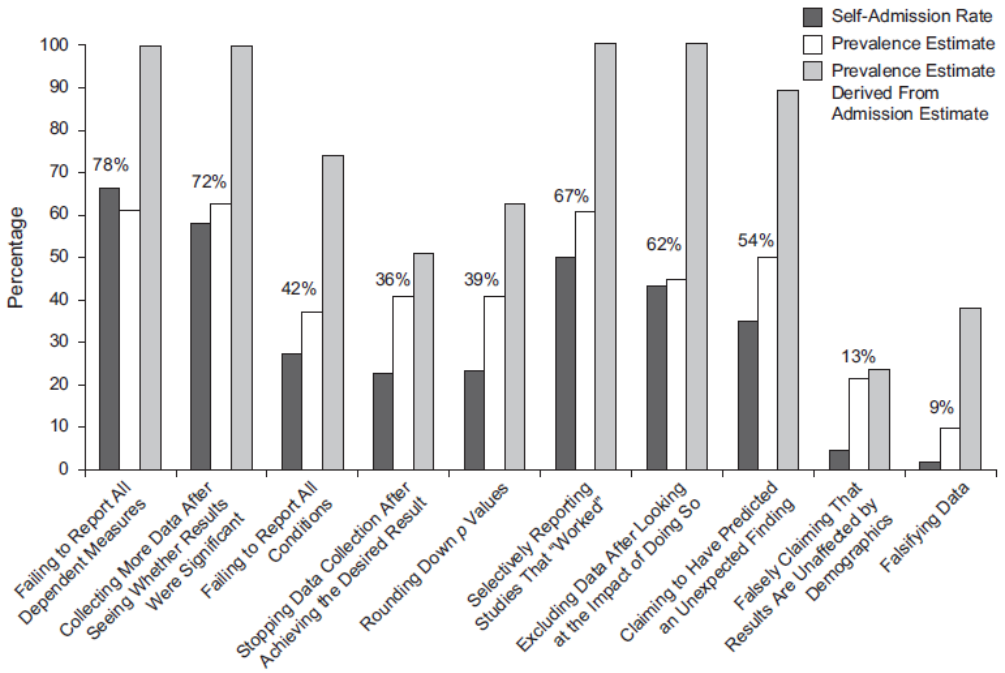


Figura 1. Risultati della condizione *Bayesian-truth-serum* (BTS) nello studio di John et al. (2012). Bayesian

Una replica di John et al. (2012) è stata eseguita da Agnoli, Wicherts, Veldkamp, Albiero, e Cubelli (2017), ma con ricercatori psicologi italiani. Le differenze principali stanno nel numero di risposte ottenute alle mail di partecipazione inviate, 277 risposte in Agnoli et al. (2017) vs 2155 risposte in John et al. (2012), ed il non utilizzo della condizione BTS nello studio-replica. Dai risultati di Agnoli et al. (2017) emerge come circa l'88% dei partecipanti ammettessero di aver utilizzato almeno una volta una delle QRP elencate. Una differenza significativa nei risultati sta nella difendibilità delle QRP. Nonostante i partecipanti ammettessero di utilizzarle, tendevano a giustificare e difendere meno il loro comportamento rispetto ai ricercatori americani dello studio di John et al. (2012). Infine, similmente ai risultati dello studio originario circa il 31% dei partecipanti riportava di

aver avuto dei dubbi sull'integrità dei loro colleghi del proprio campo di ricerca in più occasioni.

A distanza di pochi mesi viene pubblicato un articolo complementare a quello sopracitato da parte di Simmons, Nelson, e Simonsohn (2011). Se nella pubblicazione di John et al. (2012) viene indagato l'utilizzo delle QRP, Simmons et al. (2011) indagano come i diversi gradi di libertà del ricercatore, ovvero le diverse modalità con cui uno studioso può raccogliere e analizzare i dati di un esperimento, possano influire sull'ottenimento di risultati che appaiono statisticamente significativi, ma che in realtà costituiscono dei falsi positivi. Nel loro articolo vengono inizialmente descritti due esperimenti. Il fine di questi due studi, il secondo una replica del primo, era di dimostrare qualcosa di falso, ovvero che l'ascolto da parte di un partecipante di una determinata canzone potesse modificarne l'età. Entrambi gli studi sono stati divisi in due condizioni: sperimentale (canzone che avrebbe dovuto diminuire l'età del partecipante) e controllo. Dopo l'ascolto della canzone veniva chiesto ai partecipanti, 20 per condizione, di scegliere tra cinque opzioni quanto vecchi si sentissero. I risultati venivano poi categorizzati in tre livelli: basso, medio ed alto. I dati raccolti dei singoli partecipanti nelle due ricerche sopracitate sono poi stati presi in maniera casuale ed utilizzati per creare campioni indipendenti l'uno dall'altro. In totale sono stati creati 15.000 gruppi indipendenti. Ad ogni campione creato dai risultati precedentemente raccolti sono poi state applicate quattro diverse modalità con cui i dati avrebbero potuto esser stati analizzati. L'obiettivo era quello di vedere come queste tecniche di analisi andassero poi ad influire nell'ottenimento di falsi positivi. Le quattro modalità di analisi, che corrispondono a diversi gradi di libertà con cui un ricercatore può raccogliere e analizzare i dati, consistono nella flessibilità in:

- Scelta di variabili dipendenti da utilizzare nell'analisi dati (Situazione A nella Tabella 1);
- Aggiunta di partecipanti al campione dopo la raccolta dati (Situazione B nella Tabella 1);
- Utilizzo di covariate, in questo caso il genere o interazioni tra genere e trattamento (Situazione C nella Tabella 1);
- Mancato riporto di determinate condizioni sperimentali (Situazione D nella Tabella 1).

Nei risultati sono indagati tre diversi livelli di significatività ($p < .1$, $p < .05$, $p < .01$) e gli effetti dei diversi tipi di analisi sulla probabilità di ottenere falsi positivi. Prendendo in considerazione gli esiti relativi ad un *p-value* < .05 si può constatare come il solo utilizzo della prima modalità di analisi dei dati sopracitata, quasi raddoppi la probabilità di ottenere un falso positivo (da 5% a 9.5%). Questa probabilità aumenta considerevolmente quando vengono utilizzate assieme le diverse modalità di analisi dei dati. Se tutte e quattro le modalità di raccolta ed utilizzo dei dati vengono usate assieme si può arrivare ad una probabilità di ottenere un risultato falso positivo di circa 61% per $p < .05$. Di seguito è stata ricreata la tabella contenuta nell'articolo (Tabella 1) che mostra come le diverse condizioni di flessibilità possano generare diverse probabilità di ottenere falsi positivi per livelli di significatività diversa.

	Livelli di significatività		
Gradi di libertà del ricercatore	$p < .1$	$p < .05$	$p < .01$
Situazione A: due variabili dipendenti ($r = .50$)	17.8%	9.5%	2.2%
Situazione B: aggiunta di 10 osservazioni per cella	14.5%	7.7%	1.6%
Situazione C: controllare per il genere o interazioni tra genere e trattamento	21.6%	11.7%	2.7%
Situazione D: abbandono (o non abbandono) di una delle tre condizioni	23.2%	12.6%	2.8%
Combinazione delle situazioni A e B	26.0%	14.4%	3.3%
Combinazione delle situazioni A, B e C	50.9%	30.9%	8.4%
Combinazione delle situazioni A, B, C e D	81.5%	60.7%	21.5%

Tabella 1. Percentuali di probabilità dell'ottenimento di un falso positivo a seconda della modalità di raccolta e analisi dei risultati. Le percentuali sono state divise in tre livelli di significatività ($p < .1$, $p < .05$, $p < .01$).

Un altro importante studio da considerare è quello di Bakker e Wicherts (2011) sulle erronee modalità con cui vengono riportati i risultati statistici negli articoli pubblicati. Nella loro ricerca gli autori presero una serie di studi pubblicati su diverse riviste scientifiche nel 2008 sia quelle con un alto, che con un basso, fattore d'impatto. Furono presi e ricalcolati solo i test χ^2, t ed F nei casi in cui

venivano usati per la verifica dell'ipotesi nulla e non quelli associati ad altri modelli. Nei diversi studi vennero considerati sia i *p values* effettivamente associati al risultato (esempio: $p = .028$) sia quelli riportati in maniera incompleta (esempio: $p < .05$) e sono stati ricalcolati, partendo dal tipo di test e dai gradi di libertà riportati, per vedere se i nuovi risultati ottenuti fossero congruenti con quelli indicati dagli autori originali. La ricerca è stata divisa in due studi separati. Dalla prima ricerca è emerso come su 194 studi presi in considerazione, circa il 54% di quelli con *p value* riportato in maniera completa contenesse almeno un errore statistico all'interno, mentre per quelli con *p value* riportato in maniera incompleta la percentuale di studi con almeno un errore statistico era del 37%. Venne anche analizzata la quantità di *gross errors*, ovvero errori che portano un risultato ad essere statisticamente significativo quando non lo è o viceversa, all'interno dei diversi tipi di *p values*. Dei 50 *gross errors* identificati, 46 rendevano i risultati ad essere statisticamente significativi, quando in realtà non lo erano. Inoltre, è stato visto come più errori fossero presenti in studi che riportavano *p values* in maniera completa e nelle riviste a basso impatto. Il secondo studio riprende le metodologie ed obiettivi del primo, ma avendo meno articoli provenienti da riviste a basso impatto il numero di errori statistici è risultato inferiore. Un ultimo risultato importante da considerare è quello della completezza con cui i risultati venivano riportati. Infatti, il 4% dei risultati valutati nel primo studio ed il 21% di quelli valutati nel secondo erano incompleti, il che va contro le linee guida dell'*American Psychological Association* (APA) su come riportare i dati in articoli scientifici (*Publication Manual of the American Psychological Association,* VII *Edition*, 2019, p.152).

Capitolo 1
Progetti di replica nati dalla *Replication Crisis*

A seguito delle ricerche sopra citate il percorso di analisi all'interno della psicologia ha portato alla nascita di molti progetti di repliche sistematiche di studi già pubblicati. Il fine era quello di capire quanto replicabili e riproducibili sono i risultati nel campo della ricerca psicologica. Di seguito saranno descritti i principali progetti di repliche sistematiche, ma prima occorre definire e spiegare la differenza tra replica di una ricerca e riproduzione dei risultati.

1.1 Replica e riproduzione: definizioni e tipologie

Nel fornire le definizioni di riproducibilità e replicabilità di una ricerca verrà fatto riferimento alle definizioni riportate dalla *National Academies of Sciences* nel 2019.

La riproducibilità di una ricerca consiste nella capacità, da parte di ricercatori/ricercatrici, di ottenere gli stessi risultati di uno studio a cui non hanno partecipato utilizzando: dati grezzi, modalità di coding, step computazionali, modalità di analisi ed interpretazione dei risultati che caratterizzano lo studio originale. La replicabilità, invece, consiste nella capacità di rispondere allo stesso quesito indagato nello studio originale, ma utilizzando dati raccolti da un campione diverso. La differenza principale tra riproduzione e replica sta quindi nel tipo di dati e modalità di analisi che vengono utilizzati per ottenere risultati analoghi a quelli dello studio originale. La replica può quindi essere considerata come uno studio a sé stante che parte da zero e presuppone la raccolta di nuovi dati.

Un'altra differenza che intercorre tra replica e riproduzione di uno studio sta nei gradi in cui uno studio-replica può variare le modalità di raccolta ed analisi dei dati. Sono definite come repliche dirette (*direct replication*) gli studi che cercano di ricreare gli elementi cruciali (esempi: campione, procedure, metodi) dello studio originale al fine di riottenere gli effetti presenti nel nuovo studio (Zwaan et al., 2018). Sono invece definite come repliche concettuali (*conceptual replication*) gli studi dove vengono apportate delle modifiche alle procedure originali e che possono far ottenere delle differenze rispetto alla dimensione dell'effetto studiata (Zwaan et al., 2018). Quest'ultima tipologia di

repliche può avere da un solo cambiamento teorico o metodologico a numerose modifiche (LeBel et al., 2017). Le due diverse tipologie di repliche non hanno soltanto metodi di applicazione diversa, ma anche obiettivi diversi. Lo scopo principale delle repliche dirette è infatti quello di andare a vedere se testare un'ipotesi teorica con la stessa metodologia, ma dati diversi, porta agli stessi risultati. Per questo motivo questa tipologia di studi è molto comoda per trovare falsi positivi all'interno del campo di ricerca. Obiettivo delle repliche concettuali è invece quello di valutare la solidità della teoria di base al cambiare delle diverse modalità di campionamento ed analisi dei dati. Queste repliche hanno anche lo scopo di estendere e generalizzare le teorie che prendono in considerazione (Zwaan et al., 2018). Possono essere, infatti, definite come test di generalizzazione (Nosek & Errington, 2020).

Le distinzioni tra i diversi tipi di replica appena presentate (repliche dirette e repliche concettuali), cosa significa replicare uno studio e come attuarlo in maniera ottimale all'interno della ricerca scientifica sono oggi al centro di un dibattito che coinvolge numerosi/numerose ricercatori/ricercatrici. Dai risultati di uno studio di Agnoli, Fraser, Singleton Thorn, e Fidler (2021) condotto su un gruppo di ricercatori australiani (N=198) ed italiani (N=237) emerge come la maggior parte dei ricercatori dei due paesi identifichi il concetto di replica nella definizione di replica diretta. Tuttavia, non essendo questo il tema centrale di questo elaborato finale questa discussione non verrà trattata, per un approfondimento sull'argomento si veda l'articolo di Machery (2020).

Un' ultima differenza tra repliche e riproduzioni sta nelle modalità in cui vengono valutati i risultati ottenuti. Le riproduzioni possono portare a fallimenti per i seguenti due motivi: impossibilità di riproduzione delle procedure originali nell'analisi dei dati (esempio: mancanza di dati, informazioni su come sono state fatte le analisi o software di analisi originali) e incongruenza tra risultati ottenuti nella riproduzione e risultati originali (Nosek et al., 2021). La definizione di fallimento per i tentativi di replica è invece più complicata e molti autori hanno suggerito metodi diversi per la valutazione di una replica. Per esempio, Simonsohn (2015) ha suggerito il confronto tra una stima della dimensione dell'effetto originale (quando non presente) ed i risultati ottenuti nella

replica. Un altro modo è quello proposto da Camerer et.al (2018) i quali suggeriscono di valutare se la replica rifiuti l'ipotesi nulla nella stessa direzione dello studio originale. Inoltre, quando si deve prendere una decisione sul successo o meno di uno studio-replica si deve anche tenere in considerazione l'esperienza personale del ricercatore, con la quale analizzare con occhio critico i risultati ottenuti (*Open Science Collaboration*, 2015). Dalla letteratura non troviamo soltanto proposte di natura dicotomica, ma anche modalità di analisi dei risultati che utilizzano misure continue. Un esempio è quello di Verhagen e Wagenmakers (2014) che propongono un confronto Bayesiano tra la distribuzione nulla e quella a posteriori dello studio originale.

1.2 Progetti di replica successivi alla *Replication Crisis*

Ai fini di approfondire l'analisi critica della qualità della ricerca effettuata in psicologia, svariati progetti di replica furono avviati a partire dal 2014.

Le modalità con cui queste repliche vennero condotte sono di due tipi: repliche sistematiche e repliche multi-sito. Le repliche sistematiche partono dalla selezione di un campione di ricerche (gli studi originali) che saranno poi replicati da gruppi di ricercatori/ricercatrici indipendenti. Al fine di diminuire l'interferenza data dal bias di selezione degli studi, l'obiettivo è quello di cercare di replicare un numero piuttosto ampio all'interno del campione di studi originali iniziale (Nosek et al., 2021).

Le repliche multi-sito, invece, consistono nel tentativo di replicare un effetto legato ad un singolo costrutto teorico, analizzato con campioni e setting diversi. Il fine è quello di ottenere una stima della variabilità ed eterogeneità dell'effetto, aumentandone la solidità in letteratura e generalizzabilità a più contesti (Nosek et al., 2021).

Seguendo la struttura proposta da Nosek et al. (2021) nel loro articolo *Replicability, Robustness, and Reproducibility in Psychological Science*, di seguito saranno descritti, in maniera approfondita, progetti di ricerca che hanno avuto luogo a partire dal 2015. Verranno presentati prima tre progetti di replica sistematica e poi uno di replica multi-sito.

1.2.1 *Reproducibility Project: Psychology* (RPP)

Il *Reproducibility Project: Psychology* (RPP), i cui risultati sono stati pubblicati nel 2015 dalla *Open Science Collaboration* (https://osf.io/), è stato diretto da Brian Nosek ed ha avuto la partecipazione di numerosi gruppi di ricerca (277 autori). Scopo del RPP era di valutare la replicabilità di alcuni studi sperimentali nel campo della psicologia. A questo fine sono stati selezionati e replicati 100 studi, sia sperimentali che correlazionali, provenienti da diversi settori della ricerca psicologica e pubblicati su tre riviste diverse: *Psychological Science*, *Journal of Personality and Social Psychology* e *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Le modalità di assegnazione e scelta di quali ricerche replicare sono di seguito riportate. Furono scelti soltanto articoli pubblicati nel 2008 su una delle tre riviste sopra elencate. Ad ogni gruppo di ricercatori/ricercatrici che voleva partecipare al progetto, venivano proposte 20 ricerche che più si avvicinavano ai loro interessi (basate sull'ambito in cui il gruppo pubblicava). Di questi 20 studi ne poteva essere scelto solo uno. All'interno degli articoli selezionati veniva considerata come "replicabile" soltanto l'ultima ricerca descritta (esempio: se un articolo presentava tre esperimenti, solo il terzo sarebbe stato scelto per essere replicato). All'interno delle ricerche da replicare veniva scelto un elemento principale grazie al quale sarebbe stato possibile trarre le conclusioni di successo o fallimento della replica. Le statistiche più utilizzate furono: il test *t*, il test *F* e il coefficiente di correlazione. Nei risultati sono presenti confronti tra: *p values* e le dimensioni dell'effetto (dove possibile trasformate in coefficienti correlazionali per semplificare i confronti). Venne anche fatta una metanalisi che combina gli effetti delle coppie di studi (originale e replica). Infine, vi fu anche una valutazione qualitativa sul successo di replica o meno da parte dei/delle ricercatori/ricercatrici a conclusione della replica.

La scelta di quale ricerca e quale statistica utilizzare per la replica sono state decise a seguito di un confronto con gli autori originali. Il campione iniziale di ricerche da replicare era di 488 studi, dopo una selezione iniziale si ridusse a 113. Soltanto 100 di queste ricerche ottennero una replica

entro la data di scadenza imposta dai coordinatori del progetto, per mandare i materiali.

Dai risultati di questa ricerca (*Open Science Collaboration,* 2015) durata quattro anni è emerso come il 35% degli studi replicati risultasse statisticamente significativo, contro un 97% di quelli originali. Per complementarità sono stati anche analizzati gli intervalli di confidenza (*CI*) al 95% delle repliche. Questa analisi ha evidenziato come solo il 41% degli intervalli di confidenza delle repliche contenesse la dimensione dell'effetto originale. Per 22 studi che utilizzavano altre statistiche, tra cui il chi-quadro, circa il 68% degli intervalli di confidenza conteneva la dimensione dell'effetto originale. Considerando questo fattore il tasso di replicabilità generale è salito a circa il 47%. Per quel che riguarda la dimensione dell'effetto il valore medio ottenuto dalle repliche (M = 0.197, DS = 0.257) è risultato essere nettamente inferiore rispetto a quello degli studi originali (M = 0.403, DS = 0.188). Non vi furono invece ampie differenze tra i risultati degli originali e le repliche a seconda della rivista o settore di ricerca a cui le ricerche erano originariamente associate. La valutazione qualitativa dei/delle ricercatori/ricercatrici dei successi di replica o meno ha portato a risultati simili a quelli delle analisi quantitative precedentemente illustrate (39% di successo). È stata, inoltre, trovata una relazione negativa tra la probabilità di successo della replicazione ed il valore nello studio originale del *p value*. Una maggiore restrittività del *p value* originale va ad aumentare la probabilità di ottenere risultati statisticamente significativi nella replica dello studio. Infine, dai risultati è anche emerso come le ricerche che avevano ipotesi teoriche più innovative e sorprendenti fossero le più difficili da replicare. Questo risultato è in linea con il pensiero di Romero (2019), il quale identifica come una delle cause della *Replication Crisis* sia la presenza del bias di pubblicazione, che tende a favorire la pubblicazione e diffusione di ricerche con ipotesi più innovative e sorprendenti e risultati a loro sostegno.

1.2.2 *Life Outcomes of Personality Replication* (LOOPR) *Project*

Successivo al RPP il *Life Outcomes of Personality Replication* (LOOPR) *Project* ha l'obiettivo di valutare la replicabilità di 78 associazioni tra i cinque tratti di personalità identificati

dalla *Big Five* e 48 possibili esiti di vita (esempio: essere coscienziosi è associato a buone prestazioni lavorative). Il progetto è stato condotto in sei fasi.

Nella prima fase sono state scelte quali associazioni valutare. La scelta si è basata sulla rassegna in letteratura di Ozer e Benet-Martìnez (2006), che nella Tabella 1 del loro articolo raccolgono 86 associazioni tra i cinque tratti di personalità della *Big Five* e 49 esiti di vita. Partendo da questa tabella, Soto (2019) sceglie quali associazioni riprendere nel progetto LOOPR, ottenendo, infine, un campione di 78 associazioni tra i cinque tratti di personalità della *Big Five* e 48 possibili esiti di vita.

La seconda fase si concentra sulla codifica delle risorse empiriche (campione, metodo, tecniche di analisi e risultati) degli studi originali da cui sono state prese le 78 associazioni, prima da Ozer e Benet-Martìnez (2006) e poi da Soto (2019). Questa ricerca ha portato alla codifica di 38 studi originali, il cui elenco può essere trovato nell'Appendice A dei materiali supplementari della ricerca di Soto (https://osf.io/6w8qt/).

La terza fase si è concentrata sulla creazione di un protocollo di valutazione delle associazioni. Tratti di personalità ed esiti di vita sono stati valutati in due modi diversi. Per i primi è stata utilizzata la *Big Five Inventory-2* (BFI-2), mentre per gli esiti di vita è stato creato un protocollo di valutazione che riprendesse e somigliasse a quelli utilizzati negli studi originali, e che erano stati codificati nella seconda fase. Il protocollo che si è poi creato per la valutazione dei diversi esiti di vita considerati era troppo lungo ed è stato diviso in due parti e per ogni parte è stata creata una *survey* apposita. Nella *survey* 1 vi erano la BFI-2 e la prima metà del protocollo di valutazione per gli esiti di vita, mentre nella *survey* 2 vi erano presente la BFI-2 e la seconda metà del protocollo di valutazione per gli esiti di vita.

Nella quarta fase è avvenuta la raccolta dei dati. Attraverso l'utilizzo del *Qualtrics platform sample service* sono stati creati quattro gruppi di partecipanti: due composti da giovani adulti (età compresa tra i 18 e 25 anni) e due composti da adulti (età dai 18 anni in su). Ogni gruppo aveva una

numerosità campionaria minima di 1500 partecipanti, questo al fine di ottenere una potenza statistica di circa il 97% per ogni gruppo. Dopo la formazione dei gruppi sono state somministrate ad un gruppo dei giovani adulti e ad uno degli adulti la *survey* 1. Ai due gruppi rimanenti (uno dei giovani adulti e uno degli adulti) è, invece, stata somministrata la *survey* 2.

Nella quinta fase la ricerca è stata preregistrata attraverso l'utilizzo del modello di preregistrazione *Prereg Challenge* (https://osf.io/dg9m4).

Infine, nella sesta ed ultima fase i dati vennero analizzati in due set diversi. Nel primo set di analisi venne esaminato il tasso di replicabilità delle singole associazioni considerate, mentre nel secondo set di analisi i risultati delle associazioni sono stati aggregati al fine di stimare il tasso di replicabilità della letteratura tratto di personalità-esito di vita.

Due furono le ipotesi fatte da Soto e che guidarono l'analisi dei dati:

- Ipotesi 1: il tasso di replicabilità delle associazioni tratto-esito di vita considerate non sarebbe stato del 100% a causa della presenza di falsi positivi in letteratura;
- Ipotesi 2: il tasso di replicabilità delle associazioni tratto–esito di vita sarebbe risultato essere più alto di quello di altri settori della ricerca psicologica.

La seconda ipotesi porterà nei risultati ad un confronto tra il tasso di replicabilità individuato nel progetto LOOPR e quello trovato nel RPP (*Open Science Collaboration,* 2015).

Dai risultati è emerso come il tasso di replicabilità delle associazioni sia di circa l'88%. Di seguito, Figura 2, viene riportato un grafico, presente nell'articolo di Soto (2019), che confronta i tassi di successo nelle repliche tra RPP e LOOPR (sia inizialmente osservato, che poi corretto da misure più stringenti). È anche presente una colonna che rappresenta il grado di successo delle repliche a partire da un'analisi della potenza degli studi originali (solo per il progetto LOOPR). Gli intervalli di confidenza presenti nelle colonne hanno un'accuratezza del 95%.

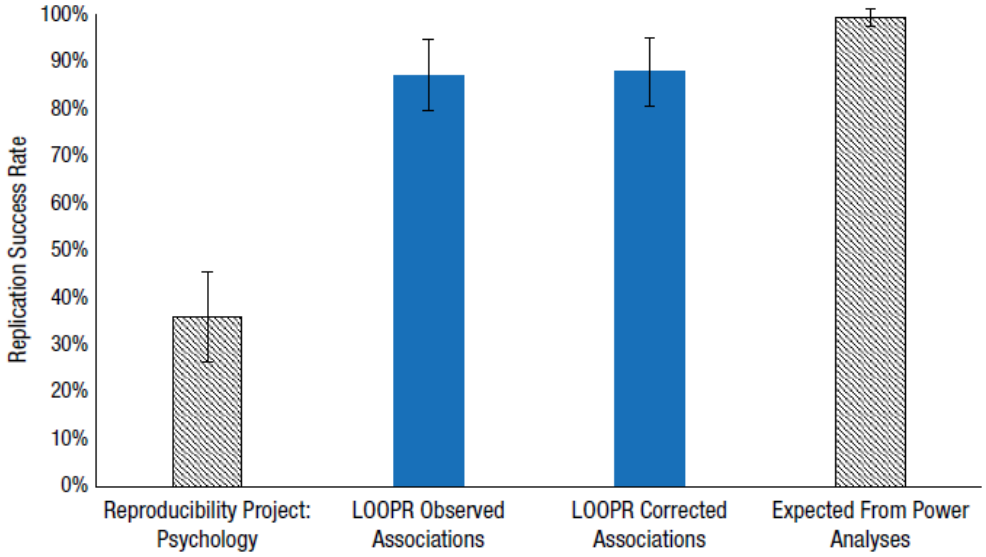


Figura 2. Confronto tra tassi di successo nelle repliche tra RPP e LOOPR. La colonna all'estrema destra rappresenta il tasso di successo nelle repliche che ci si aspettava a partire da l'analisi della potenza degli studi originali (solo progetto LOOPR).

L'analisi della potenza degli studi originali proponeva una riproducibilità dei risultati del 99.3%. Come nel caso del RPP la differenza tra presunta replicabilità e replicabilità effettiva è diversa. Circa il 71% dei risultati della replica aveva un valore di associazione tratto-esito più debole dell'originale. L'effetto replicato era, in media, più piccolo di circa il 30%. Infine, il successo della replica era positivamente predetto dall'ampiezza del campione e dalla potenza statistica dello studio originale.

1.2.3 *Social Science Replication Project* (SSRP)

Di natura più ampia è il progetto di replica sistematica svolto da Camerer et al. (2018), il cui obiettivo fu quello di valutare la replicabilità di 21 studi pubblicati tra il 2010 ed il 2011 sulle riviste scientifiche *Nature* e *Science*. Per poter essere selezionati dovevano: testare una verifica di ipotesi tramite esperimento, avere un'ipotesi chiara e sostenuta da un effetto statisticamente significativo e utilizzare un campione di partecipanti facilmente accessibile (esempio: studenti universitari). Per identificare l'effetto da replicare all'interno degli articoli selezionati, vennero utilizzati le seguenti tre fasi:

1) All'interno dell'articolo veniva selezionato il primo studio che riportasse risultati statisticamente significativi;
2) Selezionato lo studio, veniva presa tra i risultati riportati la statistica che più veniva ritenuta importante dal/dalla ricercatore/ricercatrice e associata ad un risultato statisticamente significativo;
3) Se vi fosse stata più di una statistica da poter scegliere, ne sarebbe stata selezionata una a caso.

In fase 1 le repliche selezionate avevano una potenza del 90% per ottenere 3/4 della dimensione dell'effetto originale. Se la replica avesse fallito sarebbe entrata in fase 2. Nella fase 2 veniva rifatta la replica, sempre con potenza al 90%, ma per ottenere solo la metà della dimensione dell'effetto originale. Nella seconda fase il campione veniva aumentato. La ricerca di Camerer et al. (2018) è stata preregistrata sulla piattaforma OSF (https://osf.io/pfdyw/).

Svariate modalità di analisi dei dati sono stati utilizzati per la valutazione del successo di replica o meno. Dai risultati è emerso come alla fase 1, 12 repliche su 21 (57%) ottenessero un effetto statisticamente significativo nella stessa direzione dell'originale. Considerando i risultati alla fase 2 gli studi replicati con successo salgono a 13, circa il 62%. Il valore medio della dimensione dell'effetto era di 0.249 nelle repliche, rispetto ad un valore medio di 0.460 negli studi originali. Gli autori calcolarono anche un fattore di Bayes (BF) per ogni singola replica e come distribuzione a priori venne utilizzato lo studio originale. Dall'analisi del fattore di Bayes (BF) 13 studi su 21 (circa 62%) ebbero successo come repliche. Come ultimo risultato gli autori riportano che, come nel RPP, il *p value* correla in maniera negativa con la probabilità di ottenere un successo nella replica dello studio originale. Quindi, una maggiore restrittività del *p value* originale va ad aumentare la probabilità di ottenere risultati statisticamente significativi nella replica dello studio.

1.2.4 *Many Babies Project*

La serie di ricerche intitolate *Many Babies* prende spunto dai progetti sopra citati ed ha l'obiettivo di "studiare la consistenza e replicabilità di fenomeni chiave dello sviluppo, attraverso la

coordinazione e raccolta dati tra numerosi laboratori" (Frank et al., 2017). La tipologia di repliche utilizzata è quella multi-sito. Numerosi progetti presero vita a partire dal 2017. Di seguito verrà descritto ed approfondito il primo studio di questo progetto di ricerche, il *Many Babies project* 1 (MB1), il cui compito era quello di indagare e valutare la sistematicità e replicabilità dell' *Infant-directed speech* (IDS). Tutti gli altri progetti di ricerca possono essere trovati sulla pagina web del progetto *Many Babies* (https://manybabies.github.io/).

La ricerca fu per la prima volta pubblicata dal *ManyBabies Consortium* il 16 marzo 2020 sulla rivista scientifica *Advances in Methods and Practices in Psychological Science*. Le ipotesi di seguito elencate sono state formulate facendo riferimento alla letteratura sull' IDS:

- La preferenza dei partecipanti nei confronti del IDS rispetto all'*Adult-directed speech* (ADS) sarebbe stata diversa da zero e positiva;
- Secondo Newman e Hussain (2006) all'aumentare dell'età del bambino l'interesse verso l'IDS diminuisce, aumentando l'interesse per l'ADS. Questa ipotesi è stata confrontata con un'altra formulata dai coordinatori del progetto. Secondo questa nuova ipotesi più grande diventa il bambino e più l'interesse verso l'IDS aumenterebbe, in quanto quello specifico linguaggio è associato ad una maggiore competenza da parte del bambino ed un maggior numero di interazioni positive rispetto alle competenze e numero di interazioni associate all'ADS;
- La conoscenza della lingua con cui avviene l'IDS influenza la preferenza di questo tipo di comunicazione. Ci si aspettava che i bambini più grandi la cui lingua madre era quella usata nella ricerca, mostrassero una maggiore preferenza per l'IDS confrontato con l'ADS, rispetto ai bambini la cui lingua madre non era quella utilizzata nella ricerca.

Tre furono le modalità utilizzate per la raccolta dati: *single screen central fixation, eye tracking* e *head-turn preference procedure* (HPP). Ad ogni gruppo di ricerca che partecipò al progetto venne assegnato l'utilizzo di soltanto una tecnica per la raccolta dati. La presenza di tre modalità diverse per raccogliere i dati permette, nell'analisi dei risultati, non soltanto di replicare i dati in

letteratura relativi al fenomeno studiato, ma anche di capire quale tecnica abbia una migliore sensibilità nel misurare l'effetto studiato. La lingua utilizzata per tutti gli studi fu il *North America English* (NAE).

Al progetto parteciparono 67 laboratori divisi per 18 paesi diversi e per un totale di 2329 partecipanti. Ai/alle ricercatori/ricercatrici fu dato un anno di tempo per raccogliere e mandare i dati. I bambini furono divisi in quattro gruppi per fasce di età: 3.0 – 6.0 mesi, 6.1 – 9.0 mesi, 9.1 – 12.0 mesi e 12.1 – 15.0 mesi. Ad ogni laboratorio fu assegnata una sola fascia di età. Per poter essere considerato valido ogni esperimento doveva avere almeno 16 partecipanti. Lo studio del progetto fu preregistrato (https://osf.io/gf7vh). I dati furono analizzati tramite l'utilizzo di una metanalisi (e confronto con metanalisi passate) e modelli di regressione lineare.

Dai risultati è emerso come vi sia un aumento della preferenza verso l'IDS nei bambini all'aumentare dell'età. In particolare, è stato trovato un aumento di 0.05 deviazioni standard al mese. La dimensione dell'effetto trovata dal MB1 per la preferenza dell'IDS era più piccola (d =0.35) di quella valutata in precedenti metanalisi (d =0.72). Gli autori ipotizzarono che la discrepanza tra i due effetti fosse causata dalla presenza di bias di pubblicazione in letteratura. Effetti più forti furono trovati utilizzando l'HPP come metodo di raccolta dati. Infine, i bambini preferivano l'IDS rispetto all'ADS e questo effetto era più forte per i bambini la cui lingua madre era il NAE.

Capitolo 2
Dai risultati alle procedure: la preregistrazione come nuova proposta di ricerca

Dalla crescente necessità di una ricerca più trasparente, accessibile e meno concentrata sui risultati nascono diversi modelli di preregistrazione. L'implementazione di questi nuovi modelli all'interno della ricerca psicologica è ancora in atto e presuppone, non soltanto tempi e organizzazioni diversi su come far ricerca, ma anche un cambiamento culturale da parte dei/delle ricercatori/ricercatrici. In questo capitolo, dopo una prima descrizione sull'implementazione della preregistrazione in psicologia, verranno descritti i diversi modelli, le finalità e le modalità di preregistrazione di una ricerca.

2.1 Storia della preregistrazione

La necessità di un maggior controllo sulla qualità degli studi pubblicati è stata inizialmente introdotta nella ricerca scientifica medica. Dopo pubblicazioni che sottolineavano una scarsa riproducibilità dei risultati (Prinz et al., 2011) e un'alta presenza di falsi positivi (Ioannidis, 2005) furono introdotti regolamenti finalizzati a garantire una migliore qualità delle ricerche effettuate. Dal 2007 negli USA diventò per legge obbligatoria la registrazione di tutti i risultati che coinvolgessero dei trattamenti medici o terapie poi approvate dalla *Food and Drug Administration* (*Food and Drug Administration Amendments Act*, 2007). A seguire nel 2014 sia l'Unione Europea che gli USA imposero la registrazione di tutti i *clinical trials* nella ricerca medica.

Vedendo i moti di cambiamento in medicina e i problemi evidenziati dalla *Replication Crisis*, anche nel campo della psicologia vi fu richiesta per una ricerca più trasparente basata su modelli di preregistrazione. La necessità di preregistrare non venne tanto dalle riviste, quando dagli studiosi stessi. Una prima richiesta venne da Wagenmakers et al. (2012) che sottolinearono la necessità di un metodo per distinguere tra analisi confermative ed esplorative, individuando nella preregistrazione la miglior soluzione. A distanza di un anno venne pubblicata sul quotidiano *The Guardian* una lettera

aperta, scritta da Chambers e Munafò e firmata da altri 80 ricercatori/ricercatrici, rivolta alle riviste scientifiche. La lettera spiegava come fosse necessario, per migliorare qualitativamente la ricerca scientifica, l'utilizzo di modelli di preregistrazione e la loro pubblicazione su piattaforme online prima della raccolta di dati. La richiesta degli studiosi venne ascoltata e simultaneamente sia la rivista *Cortex* che la rivista *Perspectives on Psychological Science* nel 2013 iniziarono ad accogliere e pubblicare un modello di preregistrazione chiamato *Registered Report* (RR). Con il tempo sempre più riviste iniziarono ad accettare studi preregistrati e nacque anche una rivista specializzata in studi preregistrati, *Comprehensive Results in Social Psychology*, la quale accetta e pubblica soltanto studi di questo tipo. Per stimolare la condivisione tra i/le ricercatori/ricercatrici di protocolli di pianificazione degli studi e materiali di ricerca, nel 2013 furono introdotti dall' OSF tre diversi distintivi che potevano essere affiancati al proprio articolo durante la sua pubblicazione (*Badges to Acknowledge Open Practices*, 2013). A seconda del distintivo utilizzato si poteva da subito capire se: i dati fossero pubblicamente disponibili, il materiale utilizzato fosse pubblicamente disponibile e/o se lo studio fosse stato preregistrato (con o senza le analisi dei dati fruibili per la consultazione). Dallo studio di Kidwell et al. (2016) si evince come l'introduzione dei distintivi aiuti concretamente pratiche di ricerca trasparenti e più aperte. Nella loro ricerca le/gli studiose/studiosi decisero di confrontare la quantità di articoli che condividevano dati e materiali pubblicati sulla rivista *Psychological Science*, la quale nel 2014 decise di aggiungere il sistema dei distintivi sopra descritto. Il confronto è avvenuto tra studi pubblicati prima del 2014 (pre-implementazione sistema distintivi tra il 2012 ed il 2013) e dopo il 2014 (post-implementazione sistema di distintivi tra il 2014 ed il 2015), per un totale di 838 studi considerati. Dai risultati della ricerca di Kidwell et al. (2016) emerge come tra le ricerche considerate quelle che davano la possibilità di avere accesso ai materiali usati prima del 2014 fosse di circa il 12%, mentre dopo il 2014 divenne di circa il 30%. Inoltre, vi è stato anche un aumento nella condivisione dei dati che è passata da circa il 2% prima del 2014, a circa il 22% dopo l'implementazione del sistema di distintivi.

2.2 Definizione e benefici della preregistrazione

Come definito dal *Center for Open Science* (COS) la preregistrazione consiste nella procedura di pianificazione di una ricerca e successiva registrazione su una piattaforma online (esempi: AsPredicted.org o OSF.io) prima che i dati vengano raccolti e analizzati. La possibilità di caricare il protocollo di preregistrazione permette una maggiore trasparenza delle procedure utilizzate e credibilità dei risultati ottenuti. Inoltre, la preregistrazione permette una distinzione netta tra analisi esplorative, volte alla generazione di ipotesi, e analisi confermative, volte alla verifica di ipotesi (Nosek et al., 2018).

La preregistrazione permette anche di combattere l'utilizzo delle QRP e l'*HARKing* (generazione di ipotesi dopo aver guardato i dati), in quanto le ipotesi devono esser ben specificate nel protocollo iniziale. Inoltre, la preregistrazione diminuisce i gradi di libertà del ricercatore durante l'esecuzione della ricerca. Tutti i passaggi che gli studiosi devono seguire sono descritti nel protocollo pubblicato online. Questo non vieta eventuali variazioni dalla pianificazione iniziale, ma esige una spiegazione che ne giustifica la presenza.

Essendo disponibili a tutti, i protocolli di preregistrazione permettono di combattere la presenza di bias, come quello di pubblicazione. Anche se uno studio preregistrato non verrà poi pubblicato (esempio di motivi: i risultati vanno nella direzione opposta da quella inizialmente desiderata, i risultati non confermano ipotesi già presenti in letteratura) sarà comunque possibile consultarlo online (Nosek et al., 2019). Se si sono seguiti con accuratezza i passaggi descritti nel protocollo di preregistrazione la mancata pubblicazione avviene raramente, visto e considerato che uno degli obiettivi della preregistrazione è quello di diminuire l'importanza dei risultati ottenuti e concentrarsi sulle procedure di come viene condotta la ricerca. Un altro bias, contrastato dalla preregistrazione, è quello relativo alla selettività delle analisi riportate (*reporting bias*). Le analisi che verranno compiute nella ricerca devono essere riportate nel protocollo inizialmente pubblicato e di conseguenza dovranno essere tutte presenti nel report finale, previa la sua pubblicazione (van't Veer

& Giner-Sorolla, 2016).

Infine, la preregistrazione permette di raccogliere feedback dalla comunità scientifica prima della fase di raccolta dati. In questo modo il piano di ricerca può essere implementato al fine di dare una maggiore solidità e credibilità ai risultati ottenuti.

2.3 Come preregistrare una ricerca

Il processo di preregistrazione può essere suddiviso in tre fasi:

1) Scelta del modello da utilizzare;
2) scelta del formato e scrittura del protocollo di preregistrazione;
3) scelta e caricamento del protocollo di preregistrazione sulla piattaforma online.

2.3.1 Modelli di preregistrazione

I modelli per la creazione di un protocollo di preregistrazione variano a seconda della libertà che lasciano al/alla ricercatore/ricercatrice durante la stesura del protocollo di preregistrazione e al tipo di ricerca che deve essere effettuata. In generale un modello di preregistrazione è diviso nelle seguenti parti: autori, introduzione, ipotesi, metodi di raccolta dati, analisi che verranno compiute, risultati e discussione. Le parti relative ai risultati e alla discussione vengono inserite nel protocollo di preregistrazione con l'avanzamento della ricerca. Di seguito sono elencati e descritti i principali modelli utilizzati nella ricerca psicologica.

- *AsPredicted (Walton Credibility Lab, University of Pennsylvania)*: permette la creazione di un protocollo di preregistrazione attraverso la risposta ad otto domande aperte.
- *Open-Ended Registration (OSF)*: uno dei modelli più ampi e che dà una maggiore libertà agli studiosi nella stesura. Al ricercatore viene soltanto chiesto di dare una descrizione la più accurata possibile della propria ricerca.
- *Standard Pre-Data Collection Registrations (OSF)*: simile all'*Open-Ended Registratio*n, ma con necessaria specificazione sulla raccolta ed analisi dei dati.

- *Prereg Challenge (OSF)*: in questo modello viene richiesto ai/alle ricercatori/ricercatrici di descrivere la loro ricerca attraverso la risposta a 25 domande nella maniera più accurata possibile.
- *Replication Recipe* (Brandt et al., 2013): modello di preregistrazione creato appositamente per gli studi di replica. Viene chiesto ai/alle ricercatori/ricercatrici di rispondere a 36 domande suddivise in sei categorie: la natura dell'effetto considerato, il disegno dello studio-replica, le differenze tra lo studio originale e la replica, quali analisi verranno compiute e quali criteri verranno usati per definire una replica un successo, la piattaforma su cui la replica è stata registrata e il report dei risultati con discussione dello studio-replica.
- *Registered Report Protocol Preregistration (OSF)*: questo tipo di modello porta a seguire un iter di procedimento e pubblicazione di una ricerca diverso dalle altre preregistrazioni. I RR saranno di seguito discussi ed approfonditi.
- *Qualitative Preregistration Template (OSF)*: questo modello è specifico per la ricerca qualitativa e si basa sulla risposta a 13 domande, alcune opzionali, molto specifiche.
- *Preregistration for Quantitative Research in Psychology Template* (creato dalla cooperazione tra l'*American Psychological Association*, il *British Psychological Society*, la *German Psychological Society*, il *Center for Open Science* e il *Leibniz Institute for Psychology*): consiste nel primo modello standardizzato per la ricerca quantitativa. È composto da 46 item a cui rispondere in maniera esaustiva. Non è necessario rispondere a tutti, soltanto quelli principali indicati dal modello e gli item opzionali inerenti alla propria ricerca.

La scelta del modello si deve basare sulle preferenze personali dello studioso, il settore di ricerca ed il tipo di ricerca che si sta per andare a svolgere. Nell'Appendice A viene presentato il modello *AsPredicted* come esempio (vedi Appendice A).

2.3.2 Formati per la preregistrazione

I formati con cui può essere scritto un protocollo di preregistrazione sono di diverso tipo.

I formati più usati sono:

- Google doc (versione online di Word);
- Google spreadsheet (versione online di Excel);
- compilazione di un form online;
- Jupyter Notebook (software Open Source, disponibile online);
- R Markdown (attraverso l'utilizzo di Rstudio. Nell'Appendice B vi è un esempio di come utilizzare Rstudio per creare un protocollo di preregistrazione).

2.3.3 Piattaforme di preregistrazione

Dopo aver creato il proprio protocollo di preregistrazione non resta che caricarlo sulle piattaforme online in modo che possa essere consultabile da tutti/e i/le ricercatori/ricercatrici che lo desiderano. Le tre principali piattaforme utilizzate sono: *Open Science Framework* (osf.io), *AsPredicted* (aspredicted.org) e *PreReg in Psychology* (prereg-psych.org, gestito dall'istituto di Leibniz per la psicologia).

Le prime due piattaforme sopra citate permettono la preregistrazione a seguito di una valutazione, e successiva approvazione, del protocollo inviato. La piattaforma *PreReg in Psychology* offre invece due modalità distinte per la pubblicazione del proprio protocollo: la *Repository Track* e la *Lab Track*. Nella *Repository Track* il protocollo viene prima valutato e dopo la sua approvazione riceve un DOI e viene caricato nel *PsychArchives*, dove ne sarà disponibile la consultazione. Nella *Lab Track* il protocollo viene valutato e, se approvato, sarà possibile andare a raccogliere i dati utilizzando i laboratori del *Leibniz Institute for Psychology*. Alla fine della raccolta dei dati l'elaborato verrà nuovamente valutato e pubblicato nei *PsychArchives*.

2.4 *Registered Reports*: definizioni, benefici e modalità d'uso

I *Registered Reports* (RR) sono modelli di preregistrazione che prevedono diversi passaggi e numerose revisioni del protocollo prima della sua pubblicazione. Vennero inizialmente introdotti sulla rivista *Cortex* da Christopher D. Chambers (2013). Con il passare del tempo questo modello di

preregistrazione venne adottato da sempre più riviste. Nel febbraio del 2018 le riviste che pubblicavano RR erano 91 (Hardwicke & Ioannidis, 2018). Negli ultimi 3 anni le riviste che accettano RR sono aumentate e nel giugno del 2021 le riviste che pubblicano RR sono 295 (Chambers & Tzavella, 2021).

2.4.1 Iter di scrittura e pubblicazione di un RR

La scrittura e pubblicazione di un RR può essere divisa in due fasi. Nella prima fase gli/le autori/autrici eseguono la prima stesura del loro RR attraverso l'utilizzo di un modello di preregistrazione apposito (generalmente il *Registered Report Protocol Preregistration,* OSF). Il modello prevede che al suo interno vi sia un manoscritto contenente: introduzione, ipotesi, modalità di raccolta dati e successiva analisi, analisi della potenza statistica e dati di un precedente studio pilota (se presenti). Il protocollo viene poi valutato dagli editori della rivista. Dopo la valutazione l'articolo può essere rifiutato o accettato. Se il manoscritto viene inizialmente rifiutato i revisori propongono delle modifiche al piano di esecuzione della ricerca, cercando di aumentare la solidità dei metodi e modalità di attuazione. Se, invece, la valutazione ha esito positivo la preregistrazione viene definita *In Principle Acceptance* (IPA). Il protocollo di preregistrazione viene pubblicato su una piattaforma online ed entra nella seconda fase del processo di pubblicazione. In questa seconda fase i dati vengono raccolti ed analizzati seguendo il modello originale e viene, infine, redatto il manoscritto finale. La versione finale viene inviata alla rivista e valutata dagli editori. Coloro che si occuperanno della valutazione controlleranno la corretta attuazione delle procedure descritte nella preregistrazione e, se presenti, deviazioni dal piano originale. Se la valutazione ha esito negativo, il manoscritto verrà rispedito agli autori e saranno proposte delle modifiche da apportare. Se la valutazione ha esito positivo il manoscritto viene pubblicato (Chambers, 2013). Di seguito, Figura 3, è presente un'immagine che schematizza i diversi passaggi necessari per la pubblicazione di un RR.

Figura 3. Schematizzazione dell'iter di scrittura e pubblicazione di un RR. Adattata da https://www.cos.io/initiatives/registered-reports.

2.4.2 Varianti al modello classico di RR

Con il passare del tempo sono state proposte delle varianti al modello classico di RR sopra descritto. I modelli di seguito riportati fanno riferimento a descrizioni contenute nella prestampa *The past, present and future of Registered Reports* di Chambers e Tzavella (2021).

- *Results-blind review*: questo tipo di modello sposta la prima valutazione appena dopo la raccolta dei dati. Nonostante vengano velocizzati i tempi con cui la ricerca viene effettuata, questo tipo di modello non previene la presenza di *reporting bias* e non permette modifiche alle modalità di raccolta dati suggerite dagli editori.
- *Accountable replications*: il modello classico dei RR può essere utilizzato per la pubblicazione di repliche. I momenti delle valutazioni rimangono gli stessi, ma cambiano i criteri di valutazione. Vi è una maggiore attenzione su quanto la replica proposta si avvicini nei metodi allo studio originale. Indipendentemente dai risultati ottenuti, lo studio-replica verrà pubblicato.
- *Post-publication peer review RRs*: in questo modello il manoscritto inviato nella prima fase viene immediatamente pubblicato e poi pubblicamente valutato. Se la valutazione è positiva, il manoscritto va in IPA e passa alla seconda fase dove sarà infine valutato nuovamente. Se il manoscritto viene valutato con esito negativo nella prima o seconda fase verrà comunque pubblicato sulla rivista che lo ha valutato, ma non riceverà il badge di RR.

- *Publisher-level RRs*: in questo modello le fasi di scrittura, valutazione e pubblicazione rimangono identiche a quelle del modello classico proposto da Chambers (2013). Cambiano però le riviste che valutano la ricerca nelle diverse fasi. Il manoscritto sarà prima valutato da una rivista alla fase 1, e poi valutato e pubblicato su un'altra rivista nella fase 2. Nel momento della pubblicazione sarà indicata la rivista su cui è avvenuta la prima valutazione.
- *Publisher-indipendent RRs*: i passaggi rimangono gli stessi del modello classico, ma la valutazione è indipendente dalle riviste. La valutazione del manoscritto viene eseguita dalla *Peer Community in Registered Reports*, sia nella fase 1 che nella fase 2. A seguito di una seconda valutazione positiva il manoscritto può essere pubblicato su una rivista senza ulteriori valutazioni dagli editori.

2.4.3 Benefici dei RR

I benefici portati nella ricerca da parte dei RR seguono la scia di quelli riportati dalle normali preregistrazioni, con qualche eccezione.

Come per le altre preregistrazioni i RR portano ad una maggiore presenza in letteratura dei cosiddetti "risultati nulli", ovvero risultati che non confermano le ipotesi presenti nella ricerca. In questo modo i RR diminuiscono la presenza del bias di pubblicazione nella letteratura. Lo studio di Scheel et al. (2021) descrive perfettamente le differenze dei risultati che si possono trovare tra studi RR e studi non RR. Nella loro ricerca vennero confrontati i risultati di due gruppi di ricerche pubblicate tra il 2013 ed il 2018. Nel primo gruppo erano presenti 71 RR, mentre nel secondo gruppo erano presenti 152 studi non RR. Dall'analisi dei risultati presenti nei due gruppi è emerso come circa il 96% degli articoli non RR avesse dei risultati che confermavano le ipotesi presenti negli studi. Questo risultato va confrontato con quello del gruppo dei RR, dove soltanto il 47% delle ricerche aveva dei risultati che confermavano le ipotesi descritte negli studi. Di seguito, Figura 4, viene riportato il grafico presente nell'articolo di Scheel et al. (2021) che rappresenta le percentuali di ricerche RR e non RR con risultati che confermano le ipotesi presenti negli studi. Gli intervalli di

confidenza rappresentati hanno un'accuratezza del 95%.

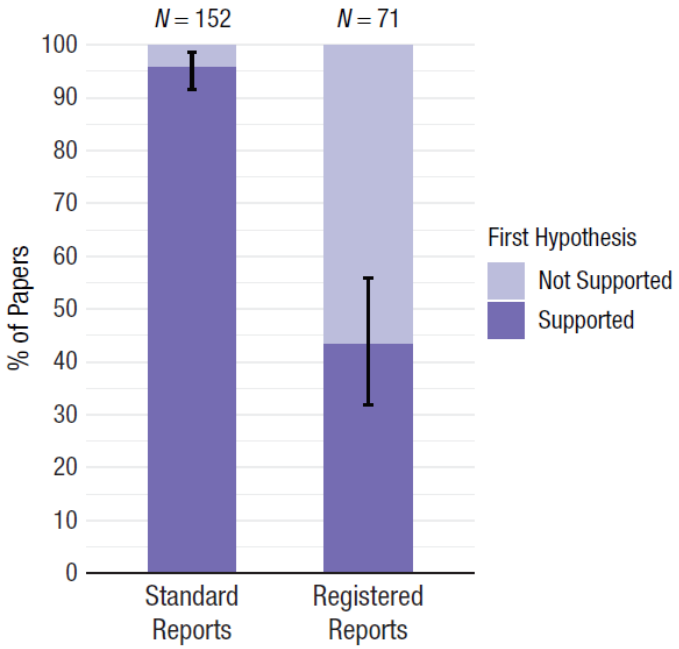


Figura 4. Percentuale di studi RR (*Registered Reports*) e non RR (*Standard Reports*) con risultati che confermano l'ipotesi presente negli studi.

L'aumento dei "risultati nulli" nei RR potrebbe essere visto come un abbassamento della qualità della ricerca. Come evidenziato dalla ricerca di Soderberg et al. (2021) questo modo di pensare non fa altro che riflettere la presenza di bias di conferma nella mente dei ricercatori, dove i risultati che confermano le ipotesi sono percepiti come positivi per il progresso della ricerca scientifica, mentre i risultati che non confermano le ipotesi sono etichettati come negativi per la letteratura. Nella loro ricerca Soderberg et al. propongono a 353 ricercatori/ricercatrici di valutare la qualità di due gruppi di ricerche. Un gruppo è composto da 29 RR, mentre l'altro gruppo è composto da 57 non RR. Gli studi presenti nei due gruppi fanno riferimento allo stesso settore di ricerca. Dai risultati emerge come i RR fossero valutati come qualitativamente superiori rispetto agli studi non RR. In particolare, le ricerche RR avevano: numerosità campionaria più ampia, maggior presenza di materiali e dati condivisi, miglior qualità delle metodologie di raccolta dati e tecniche di analisi.

L'utilizzo di RR migliora anche il clima di ricerca percepito dai/dalle ricercatori/ricercatrici. Da una ricerca di Reich et al. (2020) emerge come il processo di revisione e pianificazione nella fase

1 sia percepito come più collaborativo e finalizzato alla creazione di metodi di ricerca ottimali. Da questa ricerca emerge anche come gli autori stessi fossero più propensi ad evidenziare i limiti del loro studio nella fase di discussione dei risultati. Inoltre, dai risultati è emerso come i/le ricercatori/ricercatrici che utilizzavano i RR come protocollo di preregistrazione, riportassero un maggiore rigore nel pianificare una ricerca. Un risultato simile è emerso anche nell'esecuzione delle analisi dei dati, come riportato da un gruppo di ricerca: "abbiamo risparmiato molto tempo nell'analisi dei dati senza dover rifare dieci tipi di analisi diverse a seconda delle critiche mosse dai revisori" (Reich et al., 2020).

Un ultimo contributo dei RR è quello relativo alla possibilità di una scienza che sia aperta e trasparente. Dalla *Replication Crisis* il numero di studi volti a replicare e/o riprodurre risultati di precedenti ricerche è aumentato sempre di più. Questo ha portato anche alla messa in atto di ampi progetti di ricerca come quelli descritti nel capitolo precedente. Da un recente studio di Obels et al. (2020) si è osservato come da un campione di 62 RR (pubblicati nel periodo 2014-2018), circa il 58% avesse materiali e procedure di analisi disponibili da poter rendere possibile una riproduzione dei risultati da parte di altri ricercatori. Questo dato va confrontato con quello di Hardwicke et al. (2018), i quali ottennero un tasso di riproducibilità da 57 studi non RR (pubblicati sulla rivista *Cognition* nel periodo 2014-2017) di circa il 30%.

Capitolo 3
L'utilizzo della preregistrazione nei tentativi di replica: una ricerca sulla rivista *Psychological Science*

Con il passare degli anni l'utilizzo della preregistrazione nella ricerca psicologica è diventato sempre più comune ed accettato da diverse riviste, alcune delle quali (esempio: *Comprehensive Results in Social Psychology*) pubblicano solamente studi preregistrati. Uno dei contributi maggiori dato dalla tecnica di preregistrazione consiste, probabilmente, nel suo utilizzo nei processi di replica di ricerche già pubblicate. La preregistrazione di uno studio-replica fornisce numerosi vantaggi sia per la procedura stessa di preregistrazione, si veda l'esistenza di modelli appositi per gli studi-replica come quello di Brandt et al. (2013), sia per l'esecuzione della replica, rendendo pubblica la preregistrazione è possibile ottenere *feedback* sia dagli autori originari che da altri ricercatori/ricercatrici o nei casi in cui la replica fosse un RR (*Registered Report*) ogni passaggio sarebbe accuratamente valutato dai revisori della rivista su cui si vuole pubblicare.

Per indagare al meglio i vantaggi dati dalla preregistrazione negli studi-replica verrà di seguito riportata una ricerca eseguita sulla rivista *Psychological Science* nella quale sono stati esaminati e confrontati i risultati di 13 *Preregistered Directed Replication* con i loro rispettivi studi originari. La ricerca degli studi è avvenuta tra i mesi di aprile e giugno del 2021, mentre la loro stesura e confronto è avvenuta tra i mesi di luglio e settembre 2021. Tutti gli articoli presi in considerazione saranno di seguito descritti ed i risultati confrontati in coppia (originario-replica), dopodiché verranno tratte le conclusioni generali tra i diversi successi di replica ottenuti e l'utilità della preregistrazione.

3.1 Descrizione delle ricerche

I dati contenuti negli studi-replica di seguito riportati non sono stati raccolti prima del processo di preregistrazione. Per ogni coppia di ricerche verrà riportato prima lo studio originario e poi la replica.

Originale: Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946.

Lo studio di Fernbach et al. (2013) è centrato sull'obiettivo di studiare come i votanti (negli Stati Uniti) mantengono punti di vista molto forti (polarizzati) riguardo a politiche governative, ad esempio in ambito di riforma fiscale o in ambito di riforma sanitaria, anche quando le loro conoscenze riguardanti l'implementazione delle riforme sono molto poche. Lo studio prende in considerazione due domande di ricerca: 1) è vero che le persone hanno un'ingiustificata fiducia nella loro conoscenza rispetto a potenziali riforme (*illusion of understanding*)? 2) nel caso in cui si chieda ai votanti di spiegare come la possibile riforma sarebbe attuata, la consapevolezza della propria ignoranza, rispetto alla riforma in questione, avrebbe come conseguenza l'espressione di atteggiamenti e comportamenti di gran lunga più moderati.

Per rispondere alle domande di ricerca, un gruppo di partecipanti forniva delle stime della propria posizione riguardo a sei questioni di carattere sociopolitico prima e dopo aver fornito delle spiegazioni riguardanti le procedure di implementazione della riforma (spiegazioni definite dagli autori *mechanistic*). A questo gruppo (*within-subjects*) si affiancava un secondo gruppo (*between-subjects*) di partecipanti a cui si chiedevano le stime di sostegno alle possibili riforme solo dopo aver generato le spiegazioni di implementazione e funzionamento della riforma.

I risultati di questo primo esperimento indicano che chiedere ai votanti di spiegare come funzionerebbero le nuove riforme avrebbe l'effetto di indebolire l'ingiustificata fiducia nelle proprie conoscenze. E di conseguenza li condurrebbe a riportare atteggiamenti molto più moderati riguardanti le riforme.

Gli esperimenti 2 e 3 sono centrali per la discussione del lavoro di replica di Crawford e Ruscio (2021). Lo scopo dell'esperimento 2 era testare la differenza nel produrre un atteggiamento più moderato tra i partecipanti a cui veniva chiesto di produrre una spiegazione del funzionamento della potenziale riforma (ad esempio, fiscale o sanitaria) rispetto ai partecipanti a cui veniva solo

chiesto di enumerare le ragioni sottostanti alla riforma. I risultati indicano che enumerare le ragioni non conduce all'atteggiamento più moderato a differenza della condizione in cui ai partecipanti viene chiesto di esplicitare il funzionamento della riforma.

L'esperimento 3 aveva lo scopo di testare come l'effetto di moderazione sopra descritto avesse anche un altro effetto legato alla volontà del partecipante di supportare economicamente il gruppo che sosteneva la politica. Il metodo utilizzato è lo stesso dell'esperimento 2. L'unica differenza risiede nella parte finale nella quale ai partecipanti veniva data una piccola somma di denaro (20 centesimi) e veniva chiesto loro se avessero voluto: donarla al gruppo che sosteneva la politica, donarla al gruppo che si opponeva alla politica, tenerla per loro stessi o rifiutare di avere il compenso economico. I risultati indicano come tra coloro che inizialmente avevano una posizione estrema a favore di una politica, la condizione di spiegazione *mechanistic* abbia avuto un effetto di moderazione nel supporto economico, diminuendo la probabilità che il partecipante donasse a favore del gruppo che sosteneva la politica.

Replica: Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: Three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman's (2013) findings. *Psychological Science*, *32* (4), 611–621.

Nel loro articolo Crawford e Ruscio (2021) tentano di replicare i risultati ottenuti negli esperimenti 2 e 3 di Fernbach et al. attraverso tre *Preregistered Directed Replications,* due per l'esperimento 2 (replica 1a e 1b) ed una per l'esperimento 3 (replica 2). Seguendo le raccomandazioni di Simonsohn (2015) in tutte le repliche gli autori hanno cercato di ottenere un campione 2.5 volte più ampio di quello originale, in modo da ottenere, per i loro studi, una potenza statistica che rilevasse l'effetto originario. Questo ha portato ad un campione di 306 partecipanti per la replica 1a, 405 per la replica 1b e 343 per la replica 2.

Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza dei campioni e alla progettazione degli studi. Tutti i materiali e le analisi

statistiche compiute negli studi-replica sono disponibili su *Open Science Framework* (https://osf.io/zep2b/).

Le repliche 1a e 1b seguono lo stesso metodo usato nello studio originario, con l'unica differenza nell'aggiunta di un *attention check* dopo la domanda sul sostenimento della riforma post manipolazione. Gli autori identificavano come successo di replica l'ottenimento di risultati statisticamente significativi nella stessa direzione di quelli originali.

Come nello studio originale di Fernbach et al., dai risultati delle repliche 1a e 1b emerge come i partecipanti assegnati alla condizione *mechanistic* mostrassero una diminuzione della comprensione verso le specifiche politiche tra prima della richiesta di spiegazione (replica 1a: $M = 3.61$, $SE = 0.14$; replica 1b: $M = 3.62$, $SE = 0.12$) e dopo la spiegazione (replica 1a: $M = 3.91$, $SE = 0.14$; replica 1b: $M = 3.88$, $SE = 0.12$). Di seguito, Figura 5, viene riportata l'immagine presente nell' articolo di Crawford e Ruscio (2021) nella quale sono rappresentate le differenze tra i risultati sopra descritti dello studio originale e dei due studi replica.

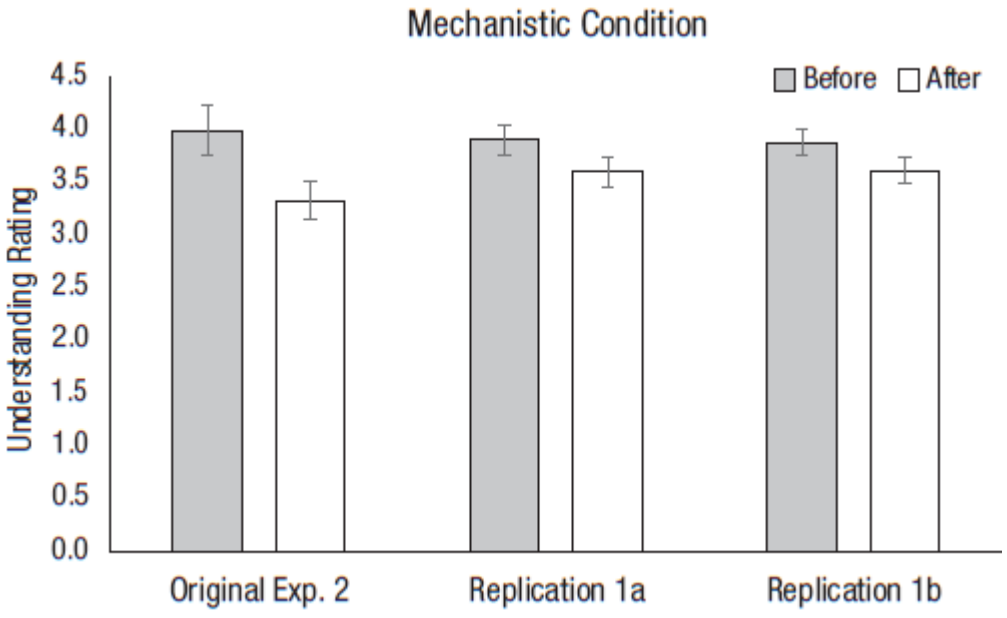


Figura 5. Differenze tra risultati originali e replica per la comprensione nella condizione *mechanistic*. Figura adattata da Crawford e Ruscio (2021).

Non è stato però trovato nei partecipanti assegnati alla condizione *mechanistic* l'effetto originario di moderazione nei confronti delle specifiche politiche tra prima della richiesta di

spiegazione (replica 1a: M = 1.40, SE = 0.08; replica 1b: M = 1.44, SE = 0.07) e dopo la spiegazione (replica 1a: M = 1.44, SE = 0.08; replica 1b: M = 1.48, SE = 0.07). Di seguito, Figura 6, viene riportata l'immagine presente nell' articolo di Crawford e Ruscio (2021) nella quale sono rappresentate le differenze tra i risultati sopra descritti dello studio originale e dei due studi replica.

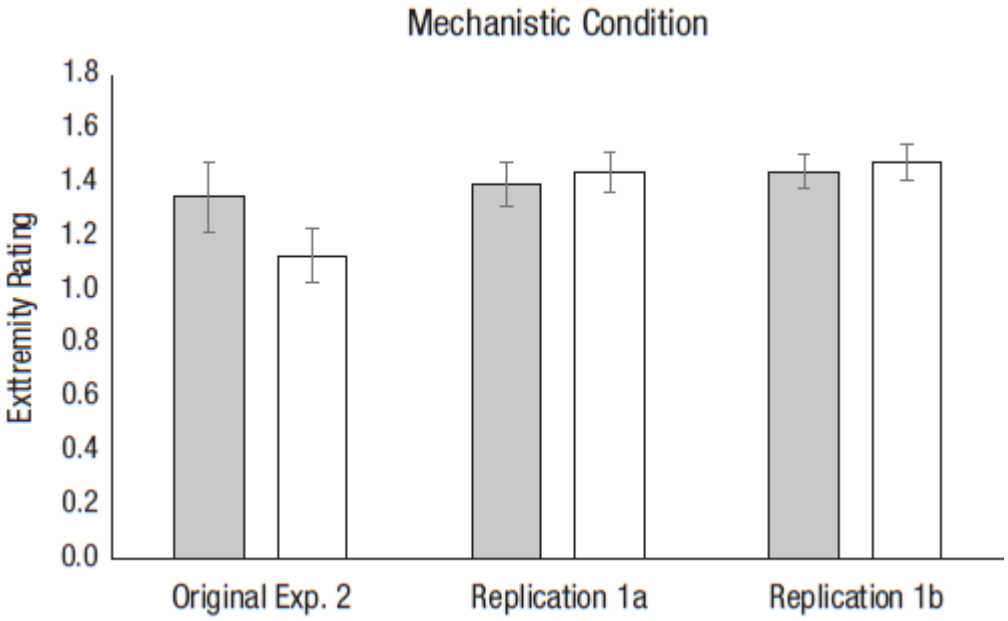


Figura 6. Differenze tra risultati originali e replica per l'effetto di moderazione nella condizione *mechanistic*. Figura adattata da Crawford e Ruscio (2021).

Come riportato da Crawford e Ruscio (2021): "*Asking participants to provide mechanistic policy explanations led them to report less policy understanding, especially relative to simply providing reasons. However, this apparent recognition of policy ignorance did not translate into political moderation for policy positions, because participants did not statistically significantly alter their issue positions after providing mechanistic explanations. We therefore failed to replicate the key finding from Fernbach et al.'s Experiments 1 and 2.*"

Nella replica 2 di Crawford e Ruscio (2021) venne utilizzato lo stesso metodo presente nell'esperimento 3 di Fernbach et al. Anche in questo caso il successo di una replica è definito dal ritrovamento di risultati statisticamente significativi nella stessa direzione di quelli originali.

Dai risultati della replica 2 emerge come la condizione *mechanistic* non abbia avuto alcun

effetto di moderazione nel supporto economico a favore del gruppo che sosteneva la politica. Coloro che inizialmente avevano una posizione estrema nei confronti di una potenziale riforma hanno deciso di supportare il gruppo, che sosteneva tale politica, tramite donazione molto più frequentemente rispetto a coloro con un'opinione più moderata, $b = 0.63$, $SE = 0.16$, Wald (1) = 16.49, $p < .001$.

Come sottolineato da Crawford e Ruscio (2021) nelle conclusioni della loro ricerca, il fallimento delle repliche non dimostra che l'effetto (indicato da Fernbach et al.) non esista. Infatti, i dati suggeriscono che se l'effetto è presente non è certamente robusto come l'avevano presentato gli autori Fernbach et al.

Originale: Kupor, D. M., Laurin, K., & Levav, J. (2015). Anticipating divine protection? Reminders of God can increase nonmoral risk taking. *Psychological Science*, *26* (4), 374-384.

Nel loro articolo Kupor et al. (2015) tentano di dimostrare, attraverso una serie di ricerche, come l'esposizione a parole associate al concetto di Dio (esempio: spirito, divino) aumentino la tendenza di una persona ad essere coinvolta in attività rischiose. Secondo gli autori l'associazione tra la concezione di Dio e il sentirsi rassicurati conduce le persone a considerare i comportamenti rischiosi come meno pericolosi e ciò, a sua volta, aumenta il potenziale coinvolgimento in attività rischiose. Tutte le attività rischiose prese in considerazione nella ricerca sono di natura non morale (esempio: paracadutismo).

Nel primo studio (1a) di Kupor et al., 61 partecipanti (reclutati da Amazon Mechanical Turk) hanno completato un compito di *priming* nel quale venivano date delle parole poi utilizzate per costruire delle frasi. Metà dei partecipanti è stata assegnata alla condizione "God" dove metà delle parole facevano riferimento al concetto di Dio; l'altra metà è stata assegnata alla condizione di controllo, dove le parole avevano un significato neutro. La misura della variabile dipendente, successiva al compito di *priming*, consisteva nel fornire la probabilità di coinvolgimento in un insieme di 40 comportamenti rischiosi su una scala che andava da 1 (molto improbabile) a 5 (molto

probabile).

Come inizialmente ipotizzato dagli autori, i partecipanti assegnati alla condizione "God" erano maggiormente propensi nell'essere coinvolti in attività rischiose (M = 2.61), rispetto ai partecipanti assegnati alla condizione di controllo (M = 2.32), t (59) = 2.21, p = .031, d = 0.574.

Nel loro secondo studio (1b) a 202 partecipanti venne inizialmente chiesto di descrivere un rischio ricreativo che avessero considerato di correre in passato. Dopodiché, venne proposto il compito di *priming* sopra descritto ed i partecipanti vennero divisi ed assegnati alle due condizioni. Alla fine del compito venne chiesto a tutti i partecipanti di rispondere alla seguente domanda: "*Qual è la probabilità che tu corra il rischio dell'attività che hai appena descritto?*" Le risposte vennero fornite con l'utilizzo di una scala che andava da 1 (estremamente improbabile) a 7 (estremamente probabile).

Dai risultati di questo secondo esperimento è emerso come i partecipanti assegnati alla condizione "God" riportassero una maggiore probabilità nel correre il rischio dell'attività da loro descritta (M = 3.38) rispetto ai partecipanti nella condizione di controllo (M = 2.77), t (200) = 2.27, p = .024, d = 0.32.

Replica: Gervais, W. M., McKee, S. E., & Malik, S. (2020). Do religious primes increase risk taking? Evidence against "anticipating divine protection" in two preregistered direct replications of Kupor, Laurin, and Levav (2015). *Psychological Science*, *31* (7), 858-864.

Nella loro ricerca Gervais et al. prendono in considerazione la replicabilità dei risultati sopra descritti con due *Preregistered Direct Replications* (PDR). In particolare fanno riferimento alle affermazioni di Field, Hoekstra, Bringmann, e van Ravenzwaaij, 2019: *Our reanalysis of [Kupor et al.'s] results, in conjunction with other methodological and theoretical criteria considerations heavily underlines this replication candidate as a promising target, reporting results that are in need of independent corroboration. We recommend a direct, or pure replication, such that the findings exactly as they are presented can be verified*.

Nella preregistrazione di Gervais et al., presente nell'appendice C, viene riportata la modalità di scelta dell'ampiezza del campione e come gli studi sono stati progettati. Tutti i materiali e le analisi condotte negli studi-replica sono disponibili su *Open Science Framework* (https://osf.io/64ct2/).

Nella replica del primo studio di Kupor et al. il campione scelto era di due volte e mezzo più ampio rispetto al campione originario. L'ampiezza del campione per la prima replica era di 566 partecipanti (rispetto ai 61 dello studio di Kupor et al.). Questo campione permette una potenza di .999 per riconoscere *l'effect size* riportato nello studio originario ($d = 0.57$).

Nella preregistrazione (https://osf.io/m28xv) gli/le autori/autrici hanno deciso di definire una replica come successo attraverso tre criteri:

- Se produce risultati statisticamente significativi nella stessa direzione dello studio originale;
- Secondo l'uso dell'approccio *small telescopes* per valutare il riconoscimento dell'effetto (vedi Simonshon, 2015);
- Secondo un'analisi del Bayes Factor (BF).

Nella replica dello studio 1a, Gervais et al. trovano i seguenti risultati:

- Non ci sono risultati statisticamente significativi, $t\ (544) = 1.618$, $p = .106$, $d = 0.24$;
- Il campione originario di 61 soggetti porta ad una potenza dell'8% associata al riconoscimento dell' *effect size* della replica ($d = 0.24$). Attraverso il criterio *small telescopes* i risultati di Gervais et al. non replicano quelli originali;
- Lo studio originario riporta un *effect size* di $d = 0.574$. Quindi il prior dell'ipotesi alternativa viene fissato a "*Normal* ~ (M =.574, SD = .267)". Questo produce un valore di BF in cui i dati supportano l'ipotesi nulla (rispetto all*'effect size* dello studio originario) per un fattore di 3.00, ciò è un supporto medio a favore dell'ipotesi nulla.

La replica dello studio 1b segue pari passo la metodologia usata nello studio originale. Il campione, però, è molto più ampio. Si tratta di 548 partecipanti. Di seguito vengono riportati i risultati dello studio replica:

- Non ci sono risultati statisticamente significativi, t (546) = -1.32, p = .187, d = - 0.1;
- L'effetto trovato nello studio replica va nella direzione opposta di quello trovato nello studio originale;
- Lo studio originario 1b riporta un *effect size* di d = 0.323. Quindi, il prior dell'ipotesi alternativa viene fissato a "*Normal* ~ (*M* =.323, *SD* = .142)". Questo produce un valore di BF in cui i dati supportano l'ipotesi nulla (rispetto all'*effect size* dello studio originario) per un fattore di 26.13. Questo valore è considerato un supporto forte a favore dell'ipotesi nulla.

Di seguito, Figura 7, viene riportata una figura presente all'interno dello studio replica di Gervais et al. (2020). L'immagine mostra le differenze, in valori ed ampiezza, degli *effect size* tra studi originali e replica.

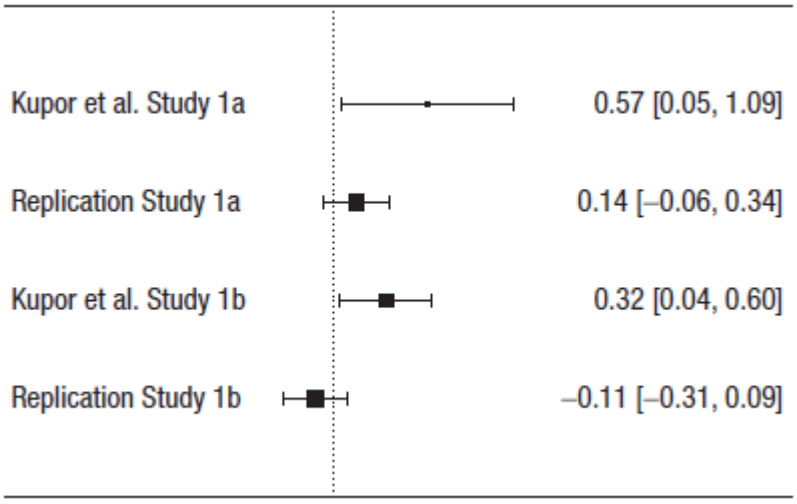


Figura 7. Differenze tra *effect size* originali e *effect size* della replica. Figura adattata da Gervais et al. (2020).

La mancata replica dei risultati di Kupor et al. è stata attribuita a due possibili motivi. La prima è l'utilizzo di una popolazione molto simile a quella di Kupor et al., che abituata agli esperimenti originali ha quasi annullato l'effetto di *priming* della manipolazione. Una seconda spiegazione è stata attribuita all'effetto stesso, non così forte come si presumeva in originale.

Originale: Rosenbaum, D., Mama, Y., & Algom, D. (2017). Stand by your Stroop: Standing up enhances selective attention and cognitive control. *Psychological Science*, *28* (12), 1864–1867.

Lo studio di Rosenbaum et al. (2017) si basa sull'assunto secondo il quale le persone debbano in ogni momento adempire a due compiti simultaneamente. Il primo consiste nel mantenimento di una determinata posizione corporea, mentre il secondo è un compito di natura cognitiva che varia da momento a momento a seconda dell'attività che si sta facendo. Sulla base di questo assunto gli autori decidono di studiare la relazione che intercorre tra una determinata posizione corporea (stare in piedi o seduti) e l'attenzione selettiva durante l'esecuzione di uno *Stroop test*. L'intento degli autori è di dimostrare che la performance di un compito cognitivo possa essere influenzata dalla postura del soggetto (in questo caso da seduti o in piedi).

Per fare ciò gli autori decidono di realizzare tre esperimenti. Nel primo esperimento (N = 17) viene chiesto ai partecipanti di eseguire uno *Stroop test* (versione con i colori) due volte, sia mentre stanno in piedi che mentre stanno seduti. A metà dei partecipanti il test viene somministrato prima nella posizione in piedi e poi da seduti, mentre per l'altra metà l'ordine delle posizioni è invertito. Metà degli stimoli colore-parola sono congruenti e metà degli stimoli colore-parola sono incongruenti.

I risultati di questo primo esperimento indicano che la media dei tempi di risposta (RT) è di 785 ms per stimoli congruenti e di 892 ms per stimoli incongruenti quando i partecipanti sono seduti. Quando, invece, i partecipanti sono in piedi la media è di 785 ms per stimoli congruenti e di 861 per stimoli incongruenti. La differenza di 31 ms tra le due posizioni a favore dell'effetto di riduzione dell'effetto Stroop nella posizione in piedi è confermata dall'interazione statisticamente significativa tra *Posture x Incongruency*: F (1, 16) = 5.70, p = .03, n_p^2 = .263. Da questo risultato i ricercatori concludono che stare in piedi permette una migliore attenzione selettiva nell'esecuzione dello *Stroop test*, rispetto ad eseguire il compito stando seduti.

Il secondo esperimento (N = 16) ha gli stessi obiettivi del primo ed utilizza lo stesso metodo. L'unica differenza si trova nel tipo di test utilizzato. In questo caso viene utilizzata la versione con le frecce dello *Stroop test*. I risultati del secondo esperimento confermano la presenza di un effetto di

Stroop minore nella posizione in piedi, rispetto a quella seduta. Le conclusioni sono le stesse del primo esperimento.

Il terzo esperimento, poi ripreso da Caron et al. (2020) per il loro studio-replica, utilizza lo stesso metodo del primo esperimento sopra descritto. L'unica differenza sta nell'aumento della numerosità campionaria (N = 50) volto ad ottenere una potenza statistica maggiore (circa 90%) rispetto a quella evidenziata nei primi due esperimenti. Questa ricerca è stata preregistrata e tutti i materiali e dati possono essere trovati su *Open Science Framework* (https://osf.io/uwzsb/). Anche dai risultati di quest'ultima ricerca emerge come l'effetto Stroop ottenuto nella condizione in piedi (M = 95.9 ms, t (49) = 14.32, p < .01, d = 2.03) sia mediamente inferiore rispetto a quello ottenuto nella posizione da seduto (M = 118.9 ms, t (49) = 16.52, p < .01, d = 2.37). Quindi, anche in questo caso, stare in piedi migliora l'attenzione selettiva nell'esecuzione dello *Stroop test* rispetto ad eseguire il compito stando seduti.

Replica: Caron, E. E., Reynolds, M. G., Ralph, B. C. W., Carriere, J. S. A., Besner, D., & Smilek, D. (2020). Does posture influence the Stroop effect? *Psychological Science*, *31* (11), 1452–1460.

Nel loro articolo Caron et al. (2020) tentano di replicare i risultati ottenuti da Rosenbaum et al. (2017), nel loro terzo esperimento, attraverso 5 studi-replica. Dei 5 studi presenti nell'articolo di Caron et al. (2020) i primi 4 consistono in repliche non preregistrate. Gli studi 1 e 2 sono stati eseguiti all'università di Waterloo (Canada), mentre gli studi 3 e 4 sono stati eseguiti all'università di Trent (Canada). L'unico esperimento replica che ha seguito fedelmente la procedura dell'esperimento 3 originario di Rosenbaum et al. (2017) è stato l'esperimento replica 1 (N =122), gli altri esperimenti presentano alcune variazioni indicate come segue. Nell'esperimento replica 2 (N = 122) venne vietato ai partecipanti di appoggiarsi alla scrivania dove si trovava il computer; nell'esperimento replica 3 (N = 99) venne chiesto di dare una risposta manuale indicando il colore corretto al posto di denominarlo a voce alta; nell'esperimento replica 4 (N = 80) venne chiesto ai partecipanti di stare su un solo piede quando dovevano eseguire il compito nella posizione in piedi. Tutti e 4 gli esperimenti avevano una

numerosità campionaria maggiore di quella dello studio originario. Nessuno delle 4 ricerche è, però, riuscita a replicare i risultati originari degli esperimenti di Rosenbaum et al.

Dopo 4 fallimenti di replica i ricercatori decisero di eseguire un quinto esperimento replica dell'esperimento 3 di Rosenbaum et al. (2017), ma questa volta preregistrandolo. Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza del campione e alla progettazione dello studio. Tutti i materiali e le analisi statistiche compiute nell studio-replica sono disponibili su *Open Science Framework* (https://osf.io/43qtn).

Il metodo di campionamento dell'esperimento replica 5 è diverso da quello delle altre ricerche riportate nello stesso articolo, in quanto gli autori decisero di utilizzare un modello Bayesiano sequenziale e quindi testare soggetti fino all'ottenimento di un BF di 5 per l'interazione *Posture x Incongruency*. L'utilizzo di questo metodo ha portato ad un campione di 61 partecipanti. Gli stimoli ed i metodi utilizzati sono identici a quelli dell'esperimento 3 originario di Rosenbaum et al. (2017).

In linea con i risultati dei 4 studi replica non preregistrati, dai risultati dell'esperimento 5 emerge come l'interazione *Posture x Incongruency* non sia statisticamente significativa, F (1, 49) = 0.13, MSE = 674.82, p = .72, η_p^2= .003, BF = 5.09, $p_{BIC}(H_0|D)$ = .196. Inoltre, l'analisi del fattore di Bayes era 5.09 volte a favore dell'interazione nulla rispetto a quella alternativa, indicando che ci sono 5.09 volte più evidenze a favore dell'ipotesi nulla rispetto a quella alternativa (Cohen's d = 0.063).

Dall'unione dei risultati delle 5 repliche, Caron et al. (2020) concludono come la posizione con cui una persona esegue uno *Stroop test* (seduto o in piedi) non influisca sulle prestazioni nel compito. Nell'identificare le cause dei fallimenti di replica i/le ricercatori/ricercatrici diedero la seguente spiegazione: "*One possibility is that there is a subtle (and, so far, unknown) factor (e.g., important differences in participant populations across institutions) that determines the presence of the effect, and we simply failed to account for that factor in our studies. Another possibility is that the influence of posture on the magnitude of the Stroop effect is unreliable and that underpowered studies sometimes yield positive (yet spurious) results*".

Originale: Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511–517.

Nel tentativo di identificare quali marcatori comportamentali possano predire dei tentativi di suicidio, Nock et al. (2010) testano una versione modificata dell'*Implicit Association Test* (IAT) su pazienti ricoverati nel reparto d'emergenza di psichiatria. Nello specifico l'obiettivo dei/delle ricercatori/ricercatrici era quello di valutare la presenza di una forte associazione cognitiva implicita tra la concezione del sé ed i concetti di morte/suicidio nei pazienti ricoverati per tentato suicidio, ed il valore predittivo di questa associazione per futuri tentativi di suicidio. Per questo motivo i partecipanti allo studio sono anche stati seguiti per i successivi sei mesi, in modo da poter monitorare i potenziali nuovi tentativi di suicidio.

Lo studio è composto da 157 partecipanti adulti (età superiore a 18 anni) che si erano presentati al reparto di emergenza psichiatrica (la città dell'ospedale non viene specificata). Secondo le analisi statistiche preliminari questo campione presentava un'adeguata potenza statistica per l'ottenimento di una dimensione dell'effetto media.

Come primo test è stata somministrata la versione alternativa dello strumento IAT per valutare la forza dell'associazione implicita tra il sé ed i concetti di morte/suicidio. Lo strumento IAT è un breve test, somministrato a computer, che utilizza i tempi di risposta dei partecipanti nella classificazione di determinati stimoli e, a seconda della velocità di risposta, valuta l'associazione implicita che il partecipante ha con questi stimoli. Nella versione utilizzata da Nock et al. (2010) sono stati utilizzati stimoli rappresentanti i concetti di morte (esempi: morte, morto, suicidio) e di vita (esempi: vivo, sopravvissuto, respirare) ed attributi che rappresentavano il sé (esempi: io, me stesso, mio) e non riferiti al sé (esempi: loro, gli altri). La forza dell'associazione implicita di ogni partecipante è stata indicizzata calcolando un punteggio D. Punteggi positivi di D associavano il sé ai concetti di morte/suicidio, mentre punteggi negativi di D associavano il sé al concetto di vita. Oltre

allo strumento IAT sono state somministrate ai partecipanti la *Self-Injurious and Behaviors Interview* (SITBI) per valutare la storia di tentati suicidi e la *Beck Scale for Suicide Ideation* per la valutazione della gravità dell'ideazione suicidaria. La valutazione successiva, dopo sei mesi, di nuovi tentativi di suicidio è avvenuta in due modi: il primo è stato quello della somministrazione della SITBI per via telefonica con i partecipanti che avevano avuto una pregressa storia di tentativi di suicidio ed il secondo è stato attraverso la valutazione di record medici per vedere se vi fossero stati altri ricoveri per tentativo di suicidio.

Dai risultati è emerso come i partecipanti ricoverati nel dipartimento di emergenza psichiatrica per un tentativo di suicidio avessero un'associazione implicita più forte con i concetti di morte/suicidio rispetto ai partecipanti ricoverati per altri motivi, t (155) = 2.46, $p < .05$. Inoltre, dai risultati è emerso come i partecipanti che avevano ottenuto punteggi positivi per l'indice D fossero più propensi a mettere in atto (nell'arco dei successivi 6 mesi) un tentativo di suicidio (circa il 31%), rispetto ai partecipanti che avevano ottenuto punteggi negativi per lo stesso indice (circa il 10%), χ^2 (1, N = 91) = 6.02, $p < .05$.

Da questi risultati Nock et al. (2010) concludono che la nuova versione del test IAT possa essere ritenuta un buon marcatore comportamentale per i tentativi di suicidio.

Replica: Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: Direct replication of the predictive validity of the suicide–implicit association Test. *Psychological Science*, *31*(1), 65–74.

Nel loro articolo Tello et al. (2020) tentano di replicare i risultati ottenuti da Nock et al. (2010) attraverso una *Preregistered Directed Replication*. La numerosità campionaria utilizzata in questa replica (N = 162) è stata scelta calcolando una potenza statistica dell'80%, volta a rilevare l'effetto originario ottenuto da Nock et al. (2010).

Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza dei campioni e alla progettazione dello studio. Tutti i materiali e le analisi

statistiche compiute nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/2mh48/).

La procedura utilizzata da Tello et al. (2020) nel loro studio-replica è la stessa dello studio originario di Nock et al. (2010). L'unica differenza sta nell'utilizzo della versione francese della *Suicide-Implicit Association Test* (S-IAT).

Dai risultati dello studio-replica emerge come i partecipanti ricoverati al reparto di emergenza psichiatrica per tentato suicidio non presentassero una più forte associazione implicita tra il sé e la morte/suicidio ($M = -0.55$, $SD = 0.35$) rispetto ai partecipanti ricoverati per altri motivi ($M = -0.54$, $SD = 0.39$), $t\ (160) = 0.14$, $p = .89$, $d = -0.02$, 95% CI = [– 0.31, 0.36]. I/le ricercatori/ricercatrici non riescono a trovare evidenze che i pazienti ammessi al reparto di emergenza di psichiatria dopo un tentato suicidio avessero un'associazione implicita statisticamente significativa tra il sé e i concetti di morte/sucidio, rispetto a coloro ammessi al reparto per altri motivi. Nonostante questo, come nello studio originario, il test è riuscito a predire futuri tentativi di suicidio dopo sei mesi dalla valutazione nei partecipanti con già una pregressa storia di tentativi di suicidio, $OR = 5.58$, 95% CI = [1.45, 21.51].

La mancata riproduzione del primo effetto identificato da Nock et al. (2010) viene imputata da Tello et al. (2020) alla troppa similarità tra i pazienti presenti nel campione. I partecipanti presenti nel gruppo di controllo avevano un numero eccessivo di storie di tentati sucidi, facendo in modo da essere clinicamente simili ai partecipanti ricoverati per un tentato suicidio recentemente. Per questo motivo, secondo Tello et al. (2020), la S-IAT non è riuscita a distinguere tra i pazienti ricoverati per un recente tentativo di suicidio e pazienti ricoverati per altri motivi psichiatrici.

Originale: Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Nella presente ricerca Pennycook et al. (2020) cercano di capire quali siano i motivi per cui le persone condividono le *fake news* e come cercare di arginare il problema. Secondo gli autori l'accettazione e condivisione delle notizie false è frutto della presenza di un bias di attenzione che si presenta durante l'utilizzo dei *social media*, i quali tendono a portare l'attenzione dell'utente non tanto sulla valutazione della veridicità di cosa sta leggendo, ma su altri fattori. Così facendo la persona si distrae più facilmente, non riuscendo a giudicare in maniera corretta quello che legge e condividendo con più facilità le *fake news*. La presente ricerca si è concentrata sullo studio delle *fake news* relative al COVID-19 ed è composta da due esperimenti. Tutti i materiali e le analisi statistiche compiute in entrambe le ricerche sono disponibili su *Open Science Framework* (https://osf.io/7d3xh/).

La prima ricerca si è concentrata sullo studio delle differenze di due meccanismi di valutazione. Il primo consiste nel valutare una notizia concentrandosi sull'accuratezza di quello che si sta leggendo, mentre il secondo consiste nella valutazione di una notizia basandosi sulla propria volontà di condividerla o meno. In questa ricerca i due processi sono stati studiati attraverso la proposta di valutazione fatta a dei partecipanti di un gruppo di titoli di giornale, con notizie vere e false, riguardanti il COVID-19.

I partecipanti (N=1000) sono stati divisi in due condizioni: condizione accuratezza e condizione condivisione. Tutti i partecipanti sono stati contattati attraverso l'utilizzo della piattaforma online LUCID. A tutti i partecipanti venivano presentate delle notizie, che potevano essere vere (15 vere) o false (15 false), relative al COVID-19. Dopodichè, per ogni notizia, veniva chiesto loro di rispondere ad un quesito. Nella condizione accuratezza la domanda era "*To the best of your knowledge, is the claim in the above headline accurate?*", mentre per la condizione condivisione la domanda era "*Would you consider sharing this story online (for example, through Facebook or Twitter?)*".

Dai risultati di questo primo esperimento emerge la presenza di un'interazione tra *headline veracity x condition*, $\beta = -0.126$, $F(1, 25586) = 42.24$, $p < .0001$. In particolare, i partecipanti

assegnati alla condizione accuratezza riuscirono a distinguere meglio le notizie vere da quelle false (Cohen's d = 0.657, 95% *confidence interval* (*CI*) = [0.477, 0.836], $F(1, 25586) = 42.24$, $p < .0001$), rispetto ai partecipanti nella condizione condivisione (d = 0.121, 95% *CI* = [0.030, 0.212], $F(1, 25586) = 6.74$, $p = .009$). Di seguito, Figura 8, è riportato il grafico presente nell'articolo di Pennycook et al. (2020) che rappresenta le differenze tra i due gruppi.

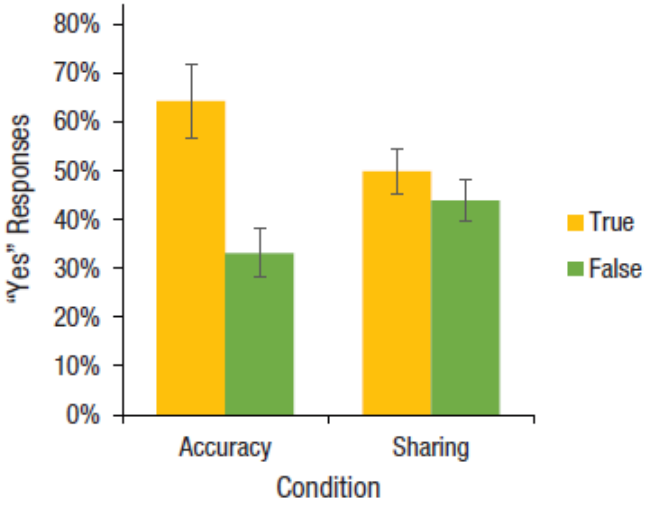


Figura 8. Percentuale di risposte "sì" per ogni combinazione di veridicità della notizia e condizione (*accuracy versus sharing*).

A partire da questi risultati Pennycook et al. (2020) concludono come la richiesta di valutazione dell'accuratezza della veridicità di una notizia aiuti nell'identificazione delle *fake news*, rispetto alla richiesta di condivisione.

Grazie ai risultati del primo esperimento, gli autori decidono di creare una seconda ricerca per valutare come la richiesta di valutazione dell'accuratezza di una notizia (non associata al COVID-19) possa portare ad una migliore capacità di distinzione tra notizie vere e false (relative al COVID-19) in un successivo compito di condivisione di quest'ultime. Per fare ciò vennero reclutati 856 partecipanti (anche in questo caso attraverso l'utilizzo della piattaforma LUCID) e furono successivamente divisi in due condizioni: trattamento e controllo. Nella condizione trattamento veniva inizialmente chiesto di valutare l'accuratezza di una notizia non associata al COVID-19, stessa domanda dell'Esperimento 1, e poi veniva proposto lo stesso compito di condivisione presente nella prima ricerca. Nella condizione

di controllo, invece, veniva soltanto presentato il compito di condivisione.

Dai risultati emerge un'interazione tra *headline veracity x condition*, β = 0.039, $F(1, 25623) = 17.88$, $p < .0001$. In particolare i risultati indicano come i partecipanti assegnati alla condizione di verifica dell'accuratezza condividessero più spesso le notizie vere rispetto a quelle false, $d = 0.142$, 95% $CI = [0.049, 0.235]$, $F(1, 25623) = 8.89$, $p = .003$, rispetto ai partecipanti assegnati alla condizione di controllo, $d = 0.050$, 95% $CI = [-0.033, 0.133]$, $F(1, 25623) = 1.41$, $p = .24$. Di seguito (vedi Figura 9) è presente il grafico riportato nella ricerca di Pennycook et al. (2020) che rappresenta le differenze tra i due gruppi.

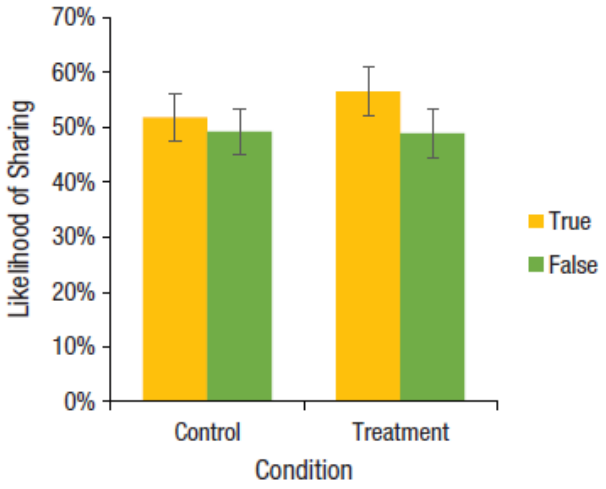


Figura 9. Percentuale di notizie condivise per ogni condizione di veridicità della notizia e condizione (*control vs treatment*).

Dai risultati ottenuti gli autori concludono come la semplice richiesta di valutazione di accuratezza di una notizia possa far diminuire la condivisione di *fake news* associate al COVID-19.

Replica: Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32* (7), 1169–1178.

Nel loro articolo Roozenbeek et al. (2021) tentano di replicare i risultati ottenuti, nel secondo studio, da Pennycook et al. attraverso una *Preregistered Direct Replication*. La scelta della numerosità campionaria (N=701) è basata sulle linee guida utilizzate nel progetto *Social Science*

Replication, che per gli studi di replica consigliano la presenza di una potenza statistica del 90%, al fine di rilevare almeno il 75% dell'effetto originario. Queste linee guida prevedono anche l'aggiunta di partecipanti per la creazione di un secondo campione con una potenza statistica del 90% che rilevi il 50% dell'effetto originario, se le prime analisi non fossero riuscite a replicare i risultati originari. In questo caso sono stati reclutati altri 822 partecipanti arrivando ad un campione di 1523 partecipanti per le seconde analisi.

Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza dei campioni e alla progettazione dello studio. Tutti i materiali e le analisi statistiche compiute nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/rkfq5/).

All'interno dello studio di Pennycook et al. (2020) i/le ricercatori/ricercatrici dello studio-replica trovano un'ambiguità tra l'ipotesi preregistrata e quello che viene riportato nei risultati. Dall'ipotesi preregistrata emerge come i ricercatori si aspettasserro che nella condizione dove si spingevano i partecipanti a valutare le notizie con l'accuratezza, questi avrebbero diminuito la condivisione di informazioni false sui *social media* riguardanti il COVID-19. Nella sezione dei risultati della ricerca, invece, gli autori riportano il risultato vedendolo sotto un altro aspetto ovvero che i partecipanti avevano condiviso più notizie vere, rispetto a quelle false. Questo non sta direttamente a significare che il numero di notizie false sia diminuito, ma che la condivisione di notizie vere è maggiore di quella delle *fake news*. Per cercare di risolvere questa ambiguità Roozenbeek et al. decidono di preregistrare due ipotesi direzionali:

1) Spingere le persone a pensare all'accuratezza di una notizia fa diminuire la probabilità che queste condividano notizie false riguardanti il COVID-19 sui *social media*;
2) Spingere le persone a pensare all'accuratezza di una notizia fa aumentare la probabilità che queste condividano notizie vere riguardanti il COVID-19 sui *social media*.

Il metodo utilizzato per la raccolta dati è lo stesso dello studio originario di Pennycook et al. (2020). Le uniche differenze stanno nell'utilizzo di notizie diverse, in quanto con il passare del tempo dall'inizio della pandemia la consapevolezza relativa al COVID-19 è aumentata, e nell'utilizzo della piattaforma *Amazon Mechanical Turk* per contattare i partecipanti. Dalle prime analisi dei risultati dello studio di replica, gli autori non riescono a trovare un'interazione tra *headline veracity x condition*, $\beta = 0.0046$, 95% *confidence interval* (*CI*) = [−0.016, 0.026], F (3, 21030) = 1.53, p =.67.

Roozenbeek et al. (2021) decidono, quindi, di proseguire con una nuova raccolta dati e successiva analisi. Dai risultati delle seconde analisi emerge un'interazione statisticamente significativa tra *headline veracity x condition*, $\beta = 0.015$, 95% *CI* = [0.0027, 0.027], F (3, 47490) = 4.52, p = .017, dimensione dell'effetto condizione trattamento: $d = -0.14$, 95% *CI* = [−0.17, −0.12]. Nonostante l'effetto fosse molto simile nel gruppo di controllo, $d = -0.10$, 95% *CI* = [−0.13, −0.078], la differenza nella condivisione era comunque 1.4 volte maggiore nella condizione di trattamento rispetto a quella di controllo. Inoltre, è stata eseguita anche un'analisi Bayesiana del t test che ha portato ad un fattore di Bayes (BF) di 1.7 volte a favore dell'ipotesi alternativa, rispetto all'ipotesi nulla, $BF_{10} = 1.705$, $M = -0.061$, 95% CI = [−0.17, 0.016], error percentage = $7.148 \times 10-6$.

Nonostante i risultati delle seconde analisi confermino parzialmente i risultati originali, Roozenbeek et al. identificano diversi motivi per il fallimento delle prime analisi dello studio di Pennycook et al. e quello parziale delle seconde analisi:

- L'esperimento è avvenuto in un momento avanzato della pandemia rispetto a quello originario, quindi la consapevolezza relativa al COVID-19 era maggiore;
- Sono state utilizzate notizie diverse;
- I dati sono stati raccolti su piattaforme diverse. *Lucid* per lo studio originale e *Amazon Mechanical Turk* per la replica;
- Altre ricerche (Branigan et al., 1999) dicono che l'effetto di *nudge (spinta)* decade dopo pochi secondi e quindi l'effetto dell'accuratezza potrebbe durare solo per le prime notizie;

- Lieve differenza nella stesura delle ipotesi.

Originale: Levinson, D. B., Smallwood, J., & Davidson R. J. (2012). The persistence of thought: Evidence for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science*, *23*(4), 375-380.

Lo studio di Levinson et al. (2012) ha l'obiettivo di studiare la relazione tra le capacità di memoria di lavoro e la presenza di pensieri intrusivi non relativi al compito che si sta svolgendo (*Task Unrelated Thought*). Partendo dalle ricerche di Christoff et al. (2009) e Stawarczyk et al. (2011) sappiamo che le zone cerebrali associate alla memoria di lavoro si attivano anche quando ci sono dei pensieri intrusivi. Inoltre, dalla ricerca di Smallwood e Schooler (2006) è emerso come nonostante i pensieri intrusivi siano un processo spontaneo e quindi con delle proprie risorse per essere attuati, questi abbiano anche una priorità rispetto ai processi utilizzati per il compito che si sta svolgendo. Da questo risultato gli autori concludono che le risorse utilizzate per la produzione di pensieri intrusivi non associati al compito che si sta svolgendo potrebbero non essere indipendenti da quelle utilizzate per lo svolgimento del compito.

Da questa serie di ricerche vengono formulati due modelli per spiegare la relazione tra pensieri intrusivi e memoria di lavoro. Il primo modello postula che i pensieri intrusivi utilizzino le risorse della memoria di lavoro per esistere. In questo caso nelle situazioni in cui le risorse mentali per la memoria di lavoro sono poche, allora lo saranno anche il numero di pensieri intrusivi presenti. Il secondo modello, invece, suggerisce che la presenza di pensieri intrusivi non sia associata alle risorse della memoria di lavoro. Secondo questo modello i pensieri intrusivi sarebbero creati in modo indipendente dalle risorse per la memoria di lavoro, il cui compito sarebbe quello di inibire la presenza di pensieri intrusivi durante l'esecuzione di un compito. In questo caso, basse risorse per la memoria di lavoro sarebbero associate ad una maggiore presenza di pensieri intrusivi durante l'esecuzione di un compito a causa della mancata inibizione di quest'ultimi.

Partendo da queste conoscenze Levinson et al. (2012) decidono di valutare i due modelli attraverso due esperimenti. Nel primo esperimento viene utilizzato un compito di ricerca visiva, mentre nel secondo, successivamente ripreso e replicato da Meier (2019), viene utilizzato un compito basato sulla respirazione.

Nel loro primo esperimento Levinson et al. (2012) chiesero a 74 partecipanti di eseguire *in primis* un compito di ricerca visiva (per la descrizione del compito si veda Forster & Lavie, 2009, esperimento 4) per valutare la presenza di pensieri intrusivi e poi l'*OSPAN task* per valutare la memoria di lavoro. Dai risultati di questa prima ricerca è emerso come i partecipanti con maggiori risorse per la memoria di lavoro riportassero anche più pensieri intrusivi durante l'esecuzione del compito di ricerca visiva.

In un tentativo di replicare i risultati sopra ottenuti i ricercatori decidono di eseguire un secondo esperimento, ma questa volta utilizzando un compito basato sulla respirazione al posto del compito di ricerca visiva. Questo compito è diviso in tre fasi: *resting baseline, breath-counting task* e *breath-awareness task*. Nell'articolo vengono riportati soltanto i risultati della terza fase. Nel *breath-awareness task*, veniva chiesto ai partecipanti (N=45, range di età: 18-65) di concentrarsi sul proprio respiro e di premere il tasto L ogni volta che espiravano. Inoltre, veniva anche chiesto di monitorare la presenza di pensieri intrusivi e di premere il tasto CONTROL ogni qual volta si pensasse a qualcosa non associato al compito. Infine, circa ogni 90 secondi (range: 60-120 secondi) sullo schermo comparivano due domande: "Giusto ora dove era la tua attenzione?" e "Quanto consapevole sei di dove si trova la tua attenzione?". Solo i risultati della prima domanda sono stati riportati nell'articolo pubblicato di Levinson et al. (2012). Per rispondere alla prima domanda veniva chiesto ai partecipanti di indicare su una scala Likert a sei punti dove le risposte si situassero partendo da "completamente sul compito" e "completamente non sul compito".

Dai risultati di questa ricerca è emerso come maggiori risorse per la memoria di lavoro fossero associate ad una maggiore presenza di pensieri intrusivi valutati attraverso la domanda, $r(40) = .33$, p

= .03. Non è stata trovata nessuna correlazione tra risorse della memoria di lavoro e consapevolezza di pensieri intrusivi tramite auto-valutazione, r (40) = −.05, p = .76.

Da questi risultati i ricercatori concludono che le risorse per la generazione di pensieri intrusivi in un compito siano le stesse per l'utilizzo della memoria di lavoro.

Replica: Meier, M. E. (2019). Is there a positive association between working memory capacity and mind wandering in a low-demanding breathing task? A preregistered replication of a study by Levinson, Smallwood, and Davidson (2012). *Psychological Science*, *30*(5), 789-797.

Nel proprio articolo Meier (2019) tenta di replicare i risultati ottenuti nel secondo esperimento di Levinson et al. (2012) attraverso una *Preregistered Directed Replication*. Nella stesura dell'articolo verranno considerate tutte e tre le parti del compito di respirazione e verrà aggiunto anche un secondo compito per la valutazione della memoria di lavoro (*symmetry span*). L'ordine di somministrazione dei compiti utilizzato è il seguente: *OSPAN task,* compiti basati sul respiro, *symmetry span*, questionari demografici.

La prima fase del compito di respirazione viene chiamata *baseline section*. In questa parte, della durata di sei minuti circa, viene chiesto ai partecipanti di guardare uno schermo nero e respirare normalmente. Ogni 90 secondi (range: 60-120 secondi) venivano proposte due domande a cui rispondere. In questo studio le domande sono le stesse riportate nello studio di Levinson et al. (2012) e, come nello studio originario, soltanto la prima verrà presa in considerazione per l'analisi dei risultati. La seconda fase del compito viene chiamata *counting section*. In questa fase veniva chiesto ai partecipanti di contare, partendo da 1, il numero di volte che espiravano ed ogni volta premere con il mignolo il tasto A sulla tastiera. Dopodiché, quando si arrivava a contare alla nona espirazione bisognava premere con il dito indice la lettera F sulla tastiera e poi ricominciare da capo. Questa parte durava circa 18 minuti e circa ogni 90 secondi (range: 60-120 secondi) venivano presentate le domande sopracitate. La terza fase, *awareness section*, del compito ha la stessa procedura di quella utilizzata da Levinson et al. (2012) nel loro studio.

La scelta della numerosità campionaria (*N*=320, range di età: 18-35) si è basata sulle raccomandazioni di Schönbrodt e Perugini (2013), per le quali un effetto correlazionale debole (circa .1) si stabilizza quando si raggiungono 250 partecipanti all'interno di un esperimento.

Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza dei campioni e alla progettazione degli studi. Tutti i materiali e le analisi statistiche compiute nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/8cwgx/).

Dai risultati dello studio-replica di Meier (2019) è emerso come i punteggi ottenuti nell'*OSPAN task*, fossero negativamente correlati con quelli ottenuti nella *awareness section* del compito basato sulla respirazione, r (251) = −.16, p = .009, 95% *CI* = [−.28, −.04]. È stato calcolato un fattore di Bayes, utilizzando i risultati ottenuti da Levinson et al. (2012) nel loro secondo esperimento come distribuzione a priori. I risultati dell'analisi bayesiana hanno portato ad un fattore (BF) di 105, indicando che i risultati ottenuti nello studio replica fossero 105 volte più a favore dell'ipotesi nulla rispetto a quella alternativa. Correlazioni negative tra i risultati nell'*OSPAN task* e fasi del compito basato sul respiro sono state trovate sia nella *baseline section*, r (251) = −.13, p = .044, 95% *CI* = [−.25, .01], BF_{10} = 0.59, e nella *counting section*, r (251) = −.19, p = .002, 95% *CI* = [−.31, −.07], BF_{10} = 8. Come si può notare utilizzando l'analisi bayesiana solo la *baseline section* ottiene risultati in linea con lo studio originario. Di seguito, Figura 10, viene riportato uno *scatterplot* che mostra l'associazione tra i punteggi ottenuti nell'*OSPAN task* e la percentuale di pensieri intrusivi in ognuna delle tre fasi del compito basato sul respiro.

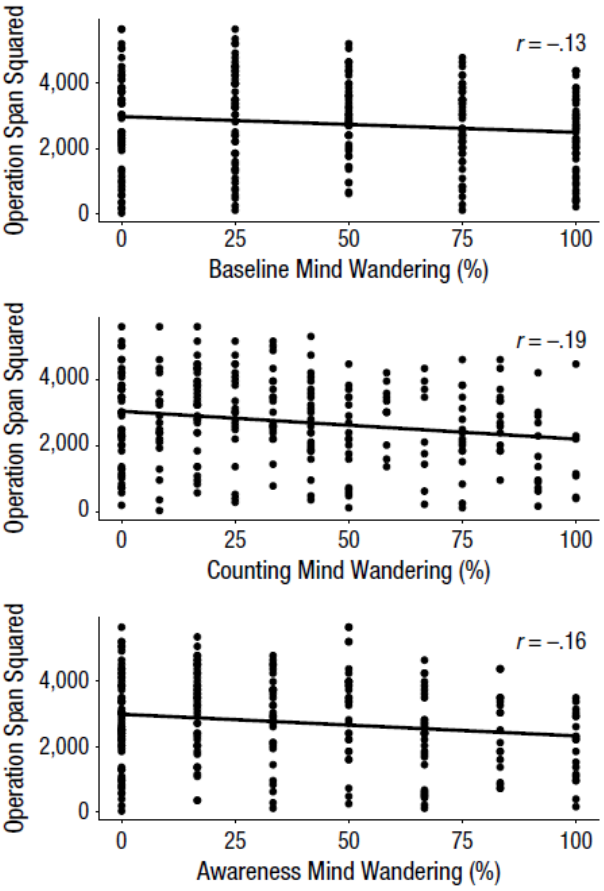


Figura 10. Scatterplot che rappresenta l'associazione tra i punteggi ottenuti nell'*OSPAN task* e la percentuale di pensieri intrusivi in ognuna delle tre fasi del compito basato sul respiro.

Inoltre, anche nei risultati di associazione tra il *symmetry span* e le tre fasi del compito basato sul respiro sono state trovate correlazioni negative. *Baseline section*, r (258) = −.14, p = .019, 95% *CI* = [−.26, −.03], BF_{10} = 1.18, *counting section*, r (258) = −.24, p < .001, 95% *CI* = [−.35, −.12], BF_{10} = 124, *awareness section*, r (258) = −.13, p = .039, 95% *CI* = [−.25, .01], BF_{10} = 0.63. Le analisi per i fattori di Bayes nella *baseliene* e *counting section* sono a favore dell'ipotesi nulla.

Similmente a Levinson et al. (2012) anche Meier (2019) non trova nessuna associazione tra risorse della memoria di lavoro e consapevolezza di pensieri intrusivi tramite auto-valutazione, sia per l'*OSPAN task*, r(251) = −.01, p = .852, 95% CI = [−.08, .20], BF_{10} = 0.08, sia per il *symmetry span*, r(258) = −.06, p = .354, 95% CI = [−.10, .18], BF_{10} = 0.12.

Dai risultati Meier (2019) conclude che i processi legati ai pensieri intrusivi non siano legati alle risorse della memoria di lavoro. La presenza di risultati divergenti, tra lo studio originario e lo

studio-replica, viene identificata da due elementi:

- Differenze nei range di età dei partecipanti nei due studi;
- Ordine in cui sono stati somministrati i compiti.

Originale: Miller, S. L., & Maner, J. K. (2011). Sick body, vigilant mind: The biological immune system activates the behavioral immune system. *Psychological Science*, *22*(12), 1467–1471.

La ricerca di Miller e Maner (2011) si basa sullo studio dell'interazione tra il sistema immunitario biologico (*Biological Immune system*), la cui attivazione promuove la distruzione di organismi patogeni che sono entrati nel corpo, e il sistema immunitario comportamentale (*Behavioural Immune system*), il cui scopo è quello di prevenire il contatto con altre persone che potrebbero essere malate. Dalla letteratura emerge come l'attivazione del sistema immunitario comportamentale attivi anche quello biologico, quindi la semplice visione di persone che sono potenzialmente malate promuove una risposta immunitaria biologica più forte (Schaller et al., 2010). Da questa base teorica Miller e Maner si chiedono se sia possible un'interazione opposta, ovvero che l'attivazione, in una persona che è stata recentemente malata, del sistema immunitario biologico possa favorire l'attivazione del sistema immunitario comportamentale. In questo modo l'individuo starà più attento alle persone che incontrerà e che mostreranno sintomi di una possibile malattia.

La presenza di questo bias di attenzione è stata indagata in due studi, di cui solo il primo sarà poi ripreso da Tybur et al. (2020) per essere replicato. Nel loro primo esperimento Miller e Maner (2011) chiedono a 96 partecipanti di completare un *dot prob task* composto da 40 fotografie di volti umani di cui 20 normali e 20 sfigurati. I volti sfigurati vennero considerati lo stimolo distraente del compito e rappresentavano persone malate. Ogni volto è comparso sullo schermo per 500 ms. Dopo il volto potevano comparire sullo schermo l'immagine di un quadrato o un cerchio ed a seconda del tipo di forma geometrica veniva chiesto al partecipante di premere un tasto. Una maggiore latenza nella risposta tra la forma geometrica e la pressione del tasto stava a significare che l'attenzione

dell'individuo era ancora sul volto che era stato precedentemente mostrato. Per valutare quanto tempo fosse passato dall'ultima volta in cui il partecipante era stato malato sono state usate sia misure continue che categoriali. Nelle valutazioni continue veniva chiesto al partecipante di indicare la propria concordanza su una scala da 1 (fortemente in disaccordo) a 7 (fortemente in accordo) sulle seguenti quattro dichiarazioni: "Negli ultimi giorni, non mi sono sentito/a molto bene", "Ultimamente mi sono sentito/a giù di corda", "Mi sono sentito/a malato/a nell'ultima settimana", "Ho avuto un raffreddore o influenza recentemente". Nella valutazione della variabile categoriale venne chiesto ai partecipanti di indicare l'ultima volta in cui avessero avuto il raffreddore o un'influenza scegliendo tra una serie di opzioni che andava da "oggi" fino a "più di un anno fa".

Dai risultati è emersa un'interazione statisticamente significativa tra l'esser stati malati recentemente e i volti sfigurati, $F(1, 92) = 9.63$, $p = .003$, $\eta_p^2 = .10$. I partecipanti che si erano ammalati nelle settimane precedenti all'esperimento (N=28) impiegavano più tempo nell'esecuzione del compito quando venivano presentati volti sfigurati (M=651 ms, DS=180) rispetto a quando venivano presentati volti normali (M = 613 ms, DS =142), $F(1, 92) = 11.06$, $p = .001$, $\eta_p^2 = .11$. I partecipanti che non si erano ammalati nelle settimane precedenti all'esperimento (N=66) non dimostravano nessun bias di attenzione, $F < 1$ (volti sfigurati: M = 618 ms, DS = 163; volti normali: M = 622 ms, DS = 182).

Da questi risultati i due ricercatori concludono che la recente attivazione del sistema immunitario biologico provochi la successiva attivazione del sistema immunitario comportamentale dopo la guarigione dalla malattia.

Replica: Tybur, J. M., Jones, B. C., DeBruine, L. M., Ackerman, J. M., & Fasolt, V. (2020). Preregistered direct replication of "sick body, vigilant mind: the biological immune system activates the behavioral immune system". *Psychological Science*, *31*(11), 1461–1469.

Nel loro studio Tybur et al. (2020) cercano di replicare i risultati del primo esperimento di Miller e Maner (2011) attraverso una *Preregistered Directed Replication.* Per la creazione del

campione è stato utilizzato il metodo suggerito da Simonsohn (2015) con il quale i/le ricercatori/ricercatrici decidono di moltiplicare di 2.5 volte la numerosità campionaria originaria ottenendo un campione di 214 partecipanti. Con il timore che i partecipanti recentemente malati fossero stati troppo pochi decisero di aumentare ancora i partecipanti all'esperimento fino ad ottenere una numerosità campionaria di 413. Questa numerosità campionaria forniva più del 99% di potenza statistica al fine di ottenere nello studio-replica un effetto d'interazione (*d*) come quello originario di 0.65.

Nella preregistrazione di Tybur et al., presente nell'appendice C, viene riportata la modalità di scelta dell'ampiezza del campione e come lo studio è stato progettato. Tutti i materiali e le analisi condotte nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/k2dbf/).

I materiali ed i metodi utilizzati sono identici a quelli dello studio originario. Nell'analisi dei risultati la numerosità campionaria è diminuita a 402 partecipanti, 151 dei quali erano stati malati di recente prima dell'esperimento e 251 non lo erano stati. Dai risultati emerge la presenza di un effetto principale tra tutti i partecipanti per il tipo di volto: le risposte erano più lente per le facce sfigurate (*M*=644 ms, *DS*=180) rispetto a quelle per i volti normali (*M* = 634 ms, *DS* = 163), $F(1, 400) = 14.96$, $\eta_p^2 = .036$, 90% CI = [.009, .078], $p < .001$. Inoltre, l'effetto principale per le persone recentemente malate non ha incontrato la soglia preregistrata ($p < .025$): tempi di risposta medi dei partecipanti recentemente malati (*M* = 661 ms, *DS* = 197); tempi di risposta medi dei partecipanti non recentemente malati (*M* = 626 ms, *DS* = 153), $F(1, 400) = 4.23$, $\eta_p^2 = .010$, 90% *CI* = [.000, .039], $p = .040$. Infine, non è stata trovata un'interazione tra l'esser stati recentemente malati e i volti sfigurati, $F(1, 400) = 1.87$, $\eta_p^2 = .005$, 90% *CI* = [.000, .027], $p = .173$. Il bias di attenzione associato all'attivazione del sistema immunitario comportamentale è stato trovato sia nei partecipanti che erano stati malati recentemente (*M* = 15.71 ms, 95% CI = [4.63, 26.79], $p = .006$), che nei partecipanti che non erano stati recentemente malati (*M* = 7.51 ms, 95% CI = [1.19, 13.83], $p = .02$).

Da questi risultati Tybur et al. (2020) concludono che non vi sia nessuna relazione tra

l'attivazione del sistema immunitario biologico e quello comportamentale, ma individuano due limiti che potrebbero aver causato che i loro risultati fossero diversi dagli originali: 1) il *dot prob task* ha numerosi limiti psicometrici e 2) il grado con cui l'avere avuto un raffreddore nelle ultime settimane possa generare una maggiore resistenza a elementi patogeni è ancora poco chiaro.

Originale: Zajonc, R. B., Heingartner, A., & Herman, E. M. (1969). Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology, 13*(2), 83–92.

Nel loro articolo Zajonc et al. (1969) valutano la teoria della facilitazione sociale attraverso lo studio del comportamento di scarafaggi (*Blatta orientalis*) in due diversi esperimenti. Secondo la teoria della facilitazione sociale la presenza di altri esemplari della stessa specie può avere un effetto di attivazione o inibizione del comportamento durante l'esecuzione di un compito, andando a migliorare o peggiorare le prestazioni nel compito stesso. In particolare, se gli stimoli che arrivano dagli altri esemplari sono appropriati alla situazione allora si vedrà nell'animale un miglioramento nelle prestazioni del compito, se, invece, gli stimoli dati dagli altri esemplari sono inappropriati si osserverà una diminuzione nelle prestazioni nel compito.

Per studiare questo effetto Zajonc et al. (1969) nel loro primo esperimento, successivamente ripreso da Halfmann et al. (2020), decidono di dividere 72 scarafaggi in due diverse condizioni (coazione e *audience)* e sottoporli a due diversi tipi di compiti: labirinto semplice e labirinto complesso. Nella condizione coazione due scarafaggi venivano messi insieme all'interno del labirinto prescelto e dovevano arrivare insieme alla fine. Nella condizione *audience*, invece, gli scarafaggi venivano messi da soli nel labirinto, ma intorno ad esso erano presenti quattro scatole con all'interno un totale di 40 scarafaggi che fungevano da "spettatori" per lo scarafaggio nel labirinto. Le scatole erano dotate di piccoli fori in modo che la presenza degli altri scarafaggi venisse riconosciuta tramite tracce olfattive. Per ogni condizione gli scarafaggi sono poi stati divisi a metà tra i due tipi di compiti.

Inoltre, per ogni condizione associata ad un compito, gli scarafaggi sono stati divisi in due gruppi: un gruppo era considerato il trattamento (l'esperimento avveniva come sopra descritto per le condizioni), mentre l'altro era il gruppo di controllo. Gli scarafaggi nel gruppo di controllo nella condizione coazione completavano i compiti da soli e non in coppia, mentre gli scarafaggi nel gruppo di controllo della condizione *audience* completavano il compito da soli e le scatole intorno al labirinto erano vuote.

Nel compito labirinto facile lo scarafaggio doveva percorrere un lungo corridoio per raggiungere la propria destinazione. Gli scarafaggi venivano posizionati nel punto di partenza e veniva poi accesa una lampadina da 150 Watt per farli muovere. L'arrivo era composto di una piccola lamiera zincata, un materiale che gli scarafaggi apprezzano. Di seguito, Figura 11, è presente un'immagine che rappresenta il labirinto semplice per la condizione *audience*. Nella condizione coazione il labirinto è identico, ma senza le scatole intorno.

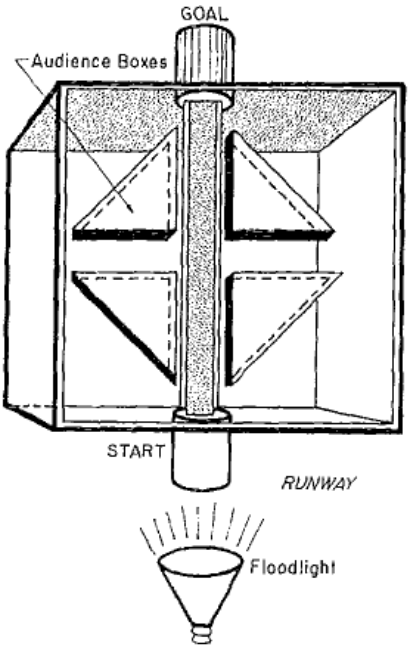


Figura 11. Rappresentazione grafica del labirinto semplice per la condizione *audience* adattata da Zajonc et al. (1969).

Nel compito labirinto complesso dopo aver percorso un breve corridoio, lo scarafaggio si trovava davanti a tre strade alternative e doveva scegliere quella corretta. I materiali utilizzati e la

procedura di partenza sono gli stessi utilizzati nel labirinto semplice. Di seguito, Figura 12, è presente un'immagine che rappresenta il labirinto complesso per la condizione *audience*. Nella condizione coazione il labirinto è identico, ma senza le scatole intorno.

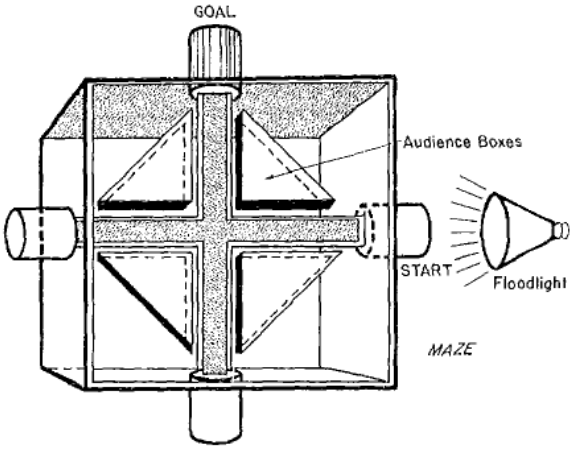


Figura 12. Rappresentazione grafica del labirinto complesso per la condizione *audience* adattata da Zajonc et al. (1969).

Da questo esperimento i/le ricercatori/ricercatrici si aspettavano che per i compiti più facili la presenza di altri scarafaggi nel labirinto (condizione coazione) o nei dintorni (condizione *audience*) migliorasse le prestazioni del soggetto e facilitasse la riuscita nel compito, mentre per i compiti più difficili la presenza di altri scarafaggi avrebbe inibito il comportamento dello scarafaggio peggiorando le prestazioni nei compiti.

Dai risultati di questo esperimento le ipotesi dei/delle ricercatori/ricercatrici vengono confermate, infatti, viene trovata un'interazione statisticamente significativa tra le condizioni e il tipo di compito (F = 7.57, df= 1/64, p < .01). In particolare, per il compito labirinto semplice gli scarafaggi che si trovavano nel gruppo di trattamento per entrambe le condizioni avevano un tempo medio di conclusione del compito inferiore (coazione: M=32.96; *audience*: M=39.30) a quello del gruppo di controllo (coazione: M=40.58; *audience*: M=62.65). Inoltre, per il compito labirinto complesso gli scarafaggi che si trovavano nel gruppo di trattamento per entrambe le condizioni avevano un tempo medio di conclusione del compito superiore (coazione: M=129.46; *audience*: M=296.64) a quello del gruppo di controllo (coazione: M=110.45; *audience*: M=221.35).

Da questi risultati i/le ricercatori/ricercatrici concludono che la teoria della facilitazione sociale sia un fenomeno esistente e presente anche nel mondo animale.

Replica: Halfmann, E., Bredehöft, J., & Häusser, J. A. (2020). Replicating roaches: A preregistered direct replication of Zajonc, Heingartner, and Herman's (1969) social-facilitation study. *Psychological Science*, *31*(3), 332–337.

Nel loro studio replica Halfmann et al. (2020) cercano di replicare i risultati ottenuti da Zajonc et al. (1969) nel loro primo esperimento attraverso una *Preregistered Directed Replication*. Nello studio di Halfmann et al. (2020) la scelta del campione di scarafaggi (N=120) è stata fatta in modo da poter determinare una dimensione dell'effetto media (.25) con una potenza statistica dell'80%.

Nella preregistrazione di Halfmann et al., presente nell'appendice C, viene riportata la modalità di scelta dell'ampiezza del campione e come lo studio è stato progettato. Tutti i materiali e le analisi condotte nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/c7t6k).

Il metodo utilizzato nello studio replica è lo stesso dell'originale, l'unica differenza sta nel tipo di scarafaggio utilizzato (*Blaberus craniifer*) e nel non utilizzo della condizione di coazione, ma soltanto quella di *audience*. I labirinti sono stati ricreati in maniera fedele a quelli originali ed adattati alle dimensioni lievemente maggiori del nuovo tipo di scarafaggi.

In linea con i risultati di Zajonc et al. (1969) i risultati dello studio replica dimostrano un tempo medio superiore nell'esecuzione del labirinto complesso (M=137.48 s, DS= 121.88) rispetto a quello del labirinto semplice (M=77.00 s, DS=76.16), $F(1, 116) = 15.45$, $p < .001$, $\eta_p^2 = .12$, BF_{10}=20.79. Inoltre, gli scarafaggi presenti nel gruppo di trattamento della condizione *audience* tendevano ad avere un tempo medio superiore nel completamento del compito, sia nel labirinto semplice che complesso, (M=164.59 s, DS= 98.84) rispetto al gruppo di controllo (M= 49.90 s, DS=77.81), $F(1, 116) = 55.58$, $p < .001$, η_p^2=.32, BF_{10} = 6.36e+7. Infine, non è stata trovata un'interazione condizione x tipo di compito come nello studio originale ed è stato calcolato un fattore

di Bayes (BF) a favore dell'ipotesi nulla, $F(1, 116) = 0.02$, $p = .882$, $\eta_p^2 = .00$, $BF_{01} = 3.88$.

Da questi risultati le/i ricercatrici/ricercatori concludono che, contrariamente allo studio originario, l'effetto d'inibizione è presente quando ci sono altri esemplari della stessa specie indipendentemente dalla difficoltà del compito.

Halfmann et al. (2020) identificano nell'utilizzo di un diverso tipo di scarafaggi il principale limite di questo studio-replica.

Originale e replica: Walmsley, J., & O'Madagain, C. (2020). The worst-motive fallacy: A negativity bias in motive attribution. *Psychological Science*, *31*(11), 1430–1438.

L'articolo di Walmsley e O'Madagain è composto da due ricerche che studiano il fenomeno del *worst-motive fallacy*, un bias cognitivo attraverso il quale gli individui si aspettano che le altre persone mettano in atto determinate azioni per perseguire motivi, da loro considerati, di natura negativa. Questi motivi generalmente portano ad effetti positivi per la persona che li persegue, ma negativi per chi gli sta intorno. I ricercatori hanno eseguito un primo studio, non preregistrato, nel quale studiano questo bias, dopodichè, sotto consiglio del loro editore, decidono di eseguire una replica preregistrata del loro primo studio, modificando il metodo utilizzato.

Nel loro primo esperimento Walmsley e O'Madagain (2020) chiedono a 323 partecipanti di completare un semplice test nel quale dovevano leggere una storia con all'interno un protagonista che aveva sia motivazioni "buone" che "cattive" per compiere una determinate azione e, succesivamente, ripondere a delle domande. In ogni storia ad un certo punto il protagonista non può più seguire entrambi i suoi obiettivi e deve quindi decidere quale dei due, quello "buono" o quello "cattivo", perseguire. Un esempio di storia utilizzata nella ricerca è di seguito riportata: "*A politician has some funding left over from her campaign, and she decides to use it to hire a computer engineer that she knows. She does this for two reasons. First, the engineer has recently lost his job and is in need of new work, and the politician wants to help him out. Second, the politician wants the engineer to send*

misleading messages to her opponent's supporters to send them to vote on the wrong day. When she describes the work to the engineer, however, the engineer says he will not do it. The politician has two further options. She could hire the unemployed engineer anyhow, to do ordinary computer maintenance work. This will help the engineer who needs income, but won't help the politician to mislead voters. Or, she could hire a computer hacker who has no problem sending misleading messages. This will help the politician to mislead voters, but will not help out the unemployed engineer." Dopo la lettura veniva chiesto al partecipante di valutare su una scala che andava da +10 (molto buono) a -10 (molto negativo) le motivazioni del protagonista e di indicare quale, secondo lui, il protagonista della storia avrebbe deciso di perseguire. La valutazione delle motivazioni è servita per valutare, anche, quanto fossero estreme le motivazioni proposte nelle storie. Se nelle diverse storie le motivazioni "buone" e "cattive" fossero state considerate più estreme delle loro controparti (a causa di un errore da parte dei ricercatori nella creazione dei materiali) allora il bias non sarebbe stato presente. Infine, veniva chiesto al partecipante quale dei due motivi presentati avrebbe voluto perseguire. Ad ogni partecipante venne assegnata una di quattro storie.

Da questa ricerca gli autori si aspettavano che se il *worst-motive fallacy* fosse stato un bias esistente allora il partecipante avrebbe detto che il protagonista della storia avrebbe scelto di perseguire la motivazione che lui (il partecipante) reputava più negativa. Inoltre, i ricercatori si aspettavano che il partecipante avrebbe poi deciso di perseguire la motivazione che, invece, riteneva essere più positiva.

Dai risultati della ricerca emerge un effetto principale dei punteggi. I partecipanti si aspettavano che il protagonista della storia avrebbe perseguito il motivo da loro valutato come più negativo, $\chi^2(1, N = 323) = -5.8005$, $p = .016$. Inoltre, è stato trovato un effetto dell'estremità dei motivi, più questi erano valutati come negativi e maggiore era la probabilità che il partecipante lo identificasse come motivo da perseguire dal protagonista della storia, $\chi^2(1, N = 323) = 8.7769$, $p = .003$. Infine, dai risultati è emerso che i partecipanti tendevano a perseguire il motivo opposto a

quello che si aspettavano avrebbe perseguito il protagonista della storia, $\chi^2(1, N = 323) = -6.607$, $p = .01$.

A seguito di questi risultati l'editore della rivista in cui l'articolo è stato pubblicato propose a Walmsley e O'Madagain (2020) di eseguire una replica preregistrata del loro studio in modo da valutare la robustezza dell'effetto. Nello studio-replica gli autori decisero di triplicare le storie presenti nella ricerca, passando dalle 4 dello studio originario a 12. Così facendo decisero anche di triplicare il numero di partecipanti per l'esperimento, passando da 323 partecipanti a 967. Le procedure utilizzate rimangono invariate rispetto allo studio originale.

Nella preregistrazione, presente in appendice C, vengono riportate le modalità di scelta relative all'ampiezza dei campioni e alla progettazione dello studio. Tutti i materiali e le analisi statistiche compiute nello studio-replica sono disponibili su *Open Science Framework* (https://osf.io/mjrpy/).

Dai risultati dello studio-replica non emerge un effetto principale dei punteggi come nello studio originario, ma è stato trovato un effetto dell'estremità dei motivi, più estremo era il motivo e maggiore era la probabilità che il partecipante lo perseguisse, $\chi^2(1, N = 967) = 39.712$, $p < .0001$. In questa replica i motivi "buoni" erano considerati più estremi di quelli "cattivi". Considerando tutte le storie i partecipanti si aspettavano che il protagonista della storia perseguisse il motivo che era considerato più negativo di quello che loro avevano scelto, $\chi^2(1, N = 967) = 97.141$, $p < .0001$. Infine, da un'analisi delle quattro storie originali è emerso come sia l'effetto principale dei punteggi, $\chi^2(1, N = 332) = 5.9072$, $p = .015$, che quello dell'estremità dei motivi, $\chi^2(1, N = 332) = 18.589$, $p < .0001$, fossero nuovamente presenti e più forti rispetto allo studio originario. Quindi, considerando soltanto le quattro vignette presenti anche nello studio originario i partecipanti si aspettavano che il protagonista della storia decidesse di perseguire il motivo che veniva valutato come più negativo e che maggiore era la negatività del motivo, maggiore sarebbe stata la probabilità che il partecipante avrebbe ritenuto che il protagonista della storia lo avrebbe perseguito.

Da questi risultati i due ricercatori identificano nelle nuove storie presenti nello studio-replica la causa della mancata replica dei risultati originali. Nonostante i risultati non siano stati replicati considerando tutte le storie, la replica è avvenuta quando vengono considerate solo le 4 storie originali. In questo caso l'effetto è anche più forte di quello originario. Da questo risultato Walmsley e O'Madagain (2020) sostengono che il bias *worst-motive fallacy* sia un fenomeno esistente.

Originale: Calogero, R. M. (2013). Objects don't object: Evidence that self-objectification disrupts women's social activism. *Psychological Science*, *24*(3), 312–318.

Con questo articolo Calogero (2013) cerca di analizzare il fenomeno dell'auto-oggettivazione nelle donne ed i suoi effetti all'interno della società odierna. L'auto-oggettivazione avviene quando una persona rivolge lo sguardo oggettivante verso sé stesso, in questo caso le donne si vedono dalla prospettiva di un osservatore esterno e mettono in atto una serie di comportamenti cronici di auto-monitoraggio. Questo fenomeno dell'auto-oggettivazione femminile è uno dei primi fenomeni che insorgono quando si vive in un contesto culturale che oggettivizza la figura della donna (Fredrickson & Roberts,1997). Un fenomeno generalmente associato a quello dell'auto-oggettivazione è quello della giustificazione del sistema sociale. La teoria della giustificazione del sistema pone che le persone sono motivate a difendere e giustificare lo status quo delle cose all'interno della società in cui vivono, anche se questo va contro i loro interessi o se si trovano in una posizione sociale svantaggiata (Jost, Banaji, & Nosek, 2004).

Da questa teoria Calogero (2013) decide di studiare il fenomeno di auto-oggettivazione femminile e le sue conseguenze: sulla messa in atto di comportamenti di attivismo sociale volti a cambiare lo status quo e favorire la parità di genere, e sulla relazione tra auto-oggettivazione e giustificazione del sistema sociale. Per fare ciò Calogero (2013) decide di eseguire due ricerche, di cui solo la prima sarà poi ripresa nell'articolo di Wilde et al. (2020).

Tre sono le ipotesi teoriche che guidano la prima ricerca:

- Alti valori di auto-oggettivazione nelle donne predicono una maggiore giustificazione del sistema sociale ed una minore messa in atto di comportamenti di attivismo sociale;
- Alti valori di giustificazione del sistema sociale nelle donne predicono la messa in atto di un minor numero di comportamenti di attivismo sociale;
- La giustificazione del sistema sociale media la relazione tra l'auto-oggettivazione la messa in atto di comportamenti di attivismo sociale.

Per valutare queste tre ipotesi Calogero chiede a 50 partecipanti (tutte donne, range di età 18-25) di completare i questionari di seguito descritti. Per la valutazione dell'auto-oggettivazione è stato utilizzato il *Self-Objectification Questionnaire*, nel quale veniva chiesto alle partecipanti di ordinare 10 attributi che corrispondono a diverse caratteristiche, 5 osservabili (esempio: attrattiva fisica) e 5 non osservabili (esempio: salute della persona), in una classifica dove nel punto più basso si trovava l'attributo che era considerato avere l'impatto più debole sul proprio concetto di sé, mentre nel punto più alto si trovava l'attributo che era considerato avere l'impatto più forte sul concetto di sé della partecipante. Il punteggio ottenuto nel questionario può variare da -25 a +25, dove punteggi più alti sono associati ad una maggiore auto-oggettivazione. La giustificazione del sistema sociale è stata valutata attraverso 8 item per ognuno dei quali la partecipante doveva fornire il suo grado di accordo su una scala che andava da 1 (fortemente in disaccordo) a 9 (fortemente in accordo). Infine, la valutazione dell'attivismo sociale è avvenuta attraverso la somministrazione di un questionario composto da 8 item che identificavano una serie di azioni, legate all'attivismo sociale per la parità di genere, che potevano essere state compiute negli ultimi 6 mesi. Per ogni comportamento veniva chiesto alle partecipanti quante volte lo avessero messo in atto tramite la risposta ad una scala di valori che andava da 1 (mai) a 7 (sempre).

Dai risultati della ricerca emerge come l'auto-oggettivazione ($M = -3.61$, $DS = 13.73$) predice l'attivismo sociale ($M = 4.06$, $DS = 1.23$), $\beta = -0.49$, $p < .001$, ed anche la giustificazione del sistema sociale ($M = 3.41$, $DS = 1.21$), $\beta = 0.52$, $p < .001$. In supporto della seconda ipotesi, dai risultati è

emerso come la giustificazione del sistema sociale predice in maniera negativa la messa in atto di comportamenti di attivismo sociale volti alla parità di genere, $\beta = -0.59$, $p < .001$. Infine, dai risultati è emerso anche l'effetto di mediatore del sistema di giustificazione sociale tra l'auto-oggettivazione ed i comportamenti di attivismo sociale.

Con questi risultati la ricercatrice conclude che alti valori di auto-oggettivazione predicono una maggiore giustificazione del sistema sociale. Inoltre, la mediazione del sistema sociale andrebbe a spiegare perché alti valori di auto-oggettivazione siano associati anche a bassi valori di attivismo sociale volto alla parità di genere. Dopo questo esperimento Calogero (2013) decise di fare un secondo esperimento con misure più stringenti dove al posto del *Self-Objectification Questionnaire* venne utilizzata una versione modificata del *Twenty Statements Test* per valutare l'auto-oggettivazione I risultati di questa seconda ricerca vanno a confermare quelli del primo studio.

Replica: De Wilde, M., Casini, A., Bernard, P., Wollast, R., Klein, O., & Demoulin, S. (2020). Two preregistered direct replications of "objects don't object: evidence that self-objectification disrupts women's social activism". *Psychological Science*, *31*(2), 214–223.

Nella loro ricerca Wilde et al. (2020) decidono di eseguire due *Preregistered Directed Replications* dell'Esperimento 1 di Calogero (2013). Sulla base dell'analisi della potenza statistica (99%) e la dimensione dell'effetto originaria vengono raccolti dati da un campione di 108 partecipanti.

Nella preregistrazione di Wilde et al., presente nell'appendice C, viene riportata la modalità di scelta dell'ampiezza del campione e come lo studio è stato progettato. Tutti i materiali e le analisi condotte nello studio-replica 1 sono disponibili su *Open Science Framework* (https://osf.io/wcy2p/).

Il metodo utilizzato è lo stesso dello studio originale di Calogero (2013). Dai risultati non emerge nessuna relazione statisticamente significativa tra l'auto-oggettivazione e la giustificazione del sistema sociale e tra l'auto-oggettivazione e la messa in atto di comportamenti di attivismo sociale volti alla parità di genere. La correlazione negativa tra la giustificazione del sistema sociale e l'attivismo sociale (r= -.34) è risultata statisticamente significativa. Inoltre, non è stato trovato nessun

effetto di mediazione da parte della giustificazione del sistema sociale, che spiegasse l'interazione tra auto-oggettivazione e la messa in atto di comportamenti di attivismo sociale.

Dopo aver ottenuto questi risultati Wilde et al. (2020) riconoscono nel loro studio due limiti principali, la numerosità campionaria e problemi legati ai criteri di esclusione utilizzati, e decidono di eseguire un secondo studio-replica della ricerca di Calogero preregistrato (2013).

Nel loro secondo studio-replica Wilde et al. (2020) decidono di utilizzare per il campionamento il metodo *small telescopes* suggerito da Simonsohn (2015) e ottengono un campione di 188 partecipanti per il loro studio. I metodi utilizzati sono gli stessi del primo studio-replica.

Nella preregistrazione di Wilde et al., presente nell'appendice C, viene riportata la modalità di scelta dell'ampiezza del campione e come lo studio è stato progettato. Tutti i materiali e le analisi condotte nello studio-replica 2 sono disponibili su *Open Science Framework* (https://osf.io/nx2sv/).

Dai risultati della ricerca non emerge nessuna relazione tra l'auto-oggettivazione e il sistema di giustificazione sociale e tra l'auto-oggettivazione e la messa in atto di comportamenti di attivismo sociale per la parità di genere. Come nel primo studio-replica viene replicata la correlazione negativa tra la giustificazione del sistema sociale e l'attivismo per la parità di genere (r= -.28). Infine, anche in questo caso non viene trovato l'effetto di mediazione della giustificazione del sistema sociale tra l'auto-oggettivazione e l'attivismo sociale.

In due studi-replica preregistrati Wilde et al. (2020) non riescono a replicare i risultati ottenuti nello studio di Calogero (2013). I/Le ricercatori/ricercatrici identificano due possibili spiegazioni per questi risultati. Il primo sta nel fatto che le relazioni tra auto-oggettivazione, giustificazione del sistema sociale e attivismo sociale possono essere spiegate da altri fattori che non sono stati considerati in questi studi. Un secondo possibile motivo per i fallimenti di replica sta nella varietà dei partecipanti utilizzati. Nello studio di Calogero (2013) tutte le partecipanti erano studentesse che provenivano da università private, mentre nelle repliche le partecipanti presentavano una maggiore eterogeneità.

3.2 Ruolo della preregistrazione negli studi esaminati e conclusioni dei confronti

Da un confronto generale dei tredici studi-replica considerati emerge come solo due di questi (Walmsley & O'Madagain, 2020; Roozenbeek et al., 2021) siano riusciti parzialmente a replicare i risultati dei propri studi originari. La presenza di numerosi fallimenti di replica non soltanto sottolinea la necessità di incentivare una maggiore produzione di questo tipo di ricerche in modo da chiarire quali effetti presenti in letteratura siano realmente affidabili o meno, ma sottolinea anche l'utilità della preregistrazione. Come precedentemente detto nel Capitolo 2, la preregistrazione in molti casi assicura la successiva pubblicazione della ricerca indipendentemente dai risultati ottenuti, a patto che si siano seguiti i passi indicati nella preregistrazione. Considerando i criteri di pubblicazione attuati dalle riviste in passato è naturale concludere che senza preregistrazione questi studi con molta difficoltà sarebbero stati pubblicati. Un altro contributo della preregistrazione sta invece nella trasparenza dei processi di progettazione dello studio-replica e condivisione ai materiali utilizzati. Infatti, tutti gli studi originari, ad eccezione di Pennycook et al. (2020), non sono stati preregistrati ed i materiali utilizzati sono di difficile accesso. Nell'ambito degli studi-replica, inoltre, la preregistrazione permette di specificare l'importanza degli effetti da replicare, quali sono le divergenze dallo studio originario e con quali criteri si è scelto il campione utilizzato per cercare di ritrovare l'effetto originario.

In aggiunta, bisogna sottolineare come la preregistrazione abbia permesso di distinguere tra analisi esplorative (esempio: Roozenbeek et al. 2021) ed analisi confermative dell'effetto che si stava cercando di replicare. In caso di fallimenti o parziale replica di effetti che si stanno cercando di replicare la presenza di analisi esplorative è fondamentale per capire i motivi del fallimento.

Infine, la presenza di studi provenienti da diverse branche della ricerca psicologica indica come la preregistrazione sia un modo di far ricerca applicabile a tutti i settori.

Concludendo, di per sé i processi di preregistrazione permettono una maggiore trasparenza nei processi di ricerca ed una maggiore affidabilità e robustezza dei risultati ottenuti permettendo al

settore di accrescere le proprie conoscenze su basi solide. Nell'ambito degli studi-replica la preregistrazione raddoppia il suo valore e importanza, in quanto permette agli studi-replica, il cui ruolo è già quello di valutare effetti e teorie presenti nella letteratura scientifica, di avere una maggiore affidabilità nei risultati ottenuti da parte della comunità scientifica.

Conclusioni

Limiti, preoccupazioni e critiche alla preregistrazione

Le pratiche di preregistrazione portano con loro anche una serie di limiti che sono stati utilizzati per criticare il metodo stesso di ricerca. La preregistrazione è quindi diventata in questi anni oggetto di dibattitto tra gli studiosi, venendo definita da alcuni/e ricercatori/ricercatrici anche come una pratica dannosa per il campo di ricerca stesso (Szollosi et al., 2020; Pham & Oh, 2021).

Uno dei principali limiti identificati sta nell'incapacità della preregistrazione di evitare il totale utilizzo delle QRP alle/ai ricercatrici/ricercatori. Un esempio di tecnica che potrebbe essere utilizzata consiste nel ri-eseguire un esperimento preregistrato molte volte e riportare, in seguito, soltanto l'esperimento con i risultati che confermano le ipotesi precedentemente registrate (Yamada, 2018). Un altro tipo di frode che potrebbe essere perpetuata, nonostante lo studio venisse preregistrato, è quella del PARKing (*pre-registering after the results are known*). Attraverso il PARKing il/la ricercatore/ricercatrice deciderebbe di preregistrare il proprio studio soltanto dopo averlo già eseguito ed avendone analizzato i risultati (Yamada, 2018).

Un altro limite evidenziato dai/dalle ricercatori/ricercatrici lo si può trovare nelle diverse tipologie di studi esistenti. Nonostante esistano numerosi modelli di preregistrazione questi non sarebbero adatti, soprattutto, per le ricerche basate sull'analisi di database con dati già esistenti. Durante una delle conferenze organizzate dalla *Society for the Improvement of Psychological Science* (SIPS) nel 2021 si è discusso di come in psicologia clinica la preregistrazione di ricerche basate sull'analisi di database vengano poco accettate, in quanto gli editori e supervisori delle stesse riviste scientifiche stesse pensano che i dati potrebbero essere già stati analizzati.

Inoltre, bisogna citare le difficoltà burocratiche associate al processo di preregistrazione. Durante numerose presentazioni nella conferenza SIPS avvenute nel mese di giugno (2021), la lamentela e preoccupazione più frequente associata alle pratiche di preregistrazione è quella relativa ai tempi eccessivamente lunghi necessari per preregistrare ed eseguire una ricerca. Anche all'interno

della comunità scientifica questo problema viene riconosciuto e la modalità di preregistrazione più colpita sarebbe quella dei *Registered Report*, per la quale sono sottolineati i tempi eccessivi per la prima revisione, la mancanza di linee guida chiare su come scrivere il manoscritto iniziale ed i continui rimbalzi tra ricercatore/ricercatrice e editore sulle modifiche da fare (Toth et al., 2020; Chambers & Tzavella, 2021).

Passando alle preoccupazioni percepite dai/dalle ricercatori/ricercatrici la più frequente è quella associata alle analisi aggiuntive o modifiche da apportare alla procedura dell'esperimento dopo la sua preregistrazione. Come descritto nel secondo studio presente nella ricerca di Toth et al. (2020) molti/e ricercatori/ricercatrici sono confusi/confuse quando devono apportare delle modifiche alla loro ricerca preregistrata: "*I had already preregistered my study and at the last moment I thought of a way to perfectionalize my experiment, but that was in contrast to some parts in my preregistration. I nevertheless changed the experiment, but now I am unsure what to do with the preregistration*". Come già evidenziato nel secondo capitolo questo tipo di situazioni non sono un pericolo, in quanto i protocolli preregistrati si possono modificare; l'importante è sempre spiegare il perché di queste modifiche, come sono state eseguite e come è, conseguentemente, cambiata la procedura.

Una seconda preoccupazione indicata dai/dalle ricercatori/ricercatrici nello studio di Toth et al. (2020) è quella relativa alla possibilità di fenomeni di *scooping*, ovvero che le proprie ipotesi e procedure per lo svolgimento di uno studio condivise tramite il caricamento *online* della propria preregistrazione possano essere rubate da altri/e ricercatori/ricercatrici. Per prevenire questo tipo di problemi molte piattaforme di preregistrazione danno la possibilità di mettere sotto embargo le proprie preregistrazioni, rendendole pubbliche soltanto a studio già avviato o concluso.

In conclusione, reputo che sia necessario evidenziare due critiche avanzate da Szollosi et al. (2020) e Pham e Oh (2021). I primi sottolineano come le ricerche preregistrate non siano necessariamente "buone" ricerche, in quanto uno studio che viene preregistrato potrebbe contenere teorie, procedure o metodi di analisi erronei per quello che si sta studiando. In risposta a questa

affermazione si può ricordare che la preregistrazione non deve essere considerata come una panacea volta a risolvere tutti i problemi in ambito di ricerca (Grand et al., 2018; Chambers, 2019; Chambers & Tzavella, 2021; Hardwicke & Wagenmakers, 2021; Soderberg et al., 2021), ma uno strumento da utilizzare per aumentare la trasparenza degli studi pubblicati. La critica avanzata da Pham e Oh (2021), invece, si va a concentrare sulla distinzione che viene fatta tra analisi esplorative e confermative. Secondo glie/le autori/autrici la preregistrazione tenderebbe a scoraggiare l'esecuzione di analisi esplorative, ovvero il tassello alla base del progresso nella ricerca scientifica. Come già precedentemente detto nel capitolo 2, la preregistrazione non vieta l'esecuzione di analisi esplorative in uno studio, ma richiede ai/alle ricercatori/ricercatrici di identificarle come tali e distinguerle da quelle confermative.

Conclusioni generali

In sintesi, se implementate in maniera corretta le tecniche di preregistrazione possono portare un buon contributo, in termini di trasparenza ed affidabilità, nella ricerca scientifica. Purtroppo, come evidenziato dalla ricerca di Hardwicke et al. (2021) su 188 articoli, pubblicati nel periodo 2014-2017, soltanto 5 (circa il 3%) erano stati preregistrati. Questo dato indica come l'implementazione vera e propria non sia ancora del tutto avvenuta, anche se dati come quelli di Montoya et al. (2021) dimostrano che il settore è pronto per questo nuovo tipo di ricerca. Nel loro studio i/le ricercatori/ricercatrici hanno eseguito un censimento, attraverso l'analisi delle linee guida per autori, delle svariate riviste scientifiche nelle scienze sociali. Dai risultati è emerso come 278 riviste (di cui 137 di psicologia) avessero iniziato a pubblicare anche *Registered Reports* nel periodo che va tra il 2013 ed il 2020.

Un modo per implementare l'utilizzo della preregistrazione potrebbe essere associato all'utilizzo di fondi per ricerche che vengono preregistrate, esempi si possono trovare nell'editoriale di Munafò (2017) dove viene indicato un processo per incentivare economicamente la pubblicazione di *Registered Reports* o anche in eventi come la *Preregistration Challenge* promossa dal *Center for*

Open Science, dove a seguito della pubblicazione di uno studio preregistrato veniva dato un premio di circa 1000 dollari ai ricercatori.

Bibliografia

- Agnoli, F., Fraser, H., Singleton Thorn, F., & Fidler, F. (2021). Australian and Italian Psychologists' View of Replication. *Advances in Methods and Practices in Psychological Science*, *4*(3),1-15, https://doi.org/10.1177/25152459211039218.
- 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, https://doi.org/10.1371/journal.pone.0172792.
- American Psychological Association (2019). Publication Manual of the American Psychological Association, VII *Edition*.
- Bakker, M., & Wicherts, J. M., (2011). The (mis)reporting of statistical results in psychology journals. *Behaviour Research Methods*, 43, 666-678, https://doi.org/10.3758/s13428-011-0089-5.
- Bosnjak, M., Fiebach, C., Mellor, D. T., … Sokol-Chang, R. (2021, May 13). A template for preregistration of quantitative research in psychology: Report of the Joint Psychological Societies Preregistration Task Force. https://doi.org/10.31234/osf.io/d7m5r.
- Brandt, M. J., Ijzerman, H., Dijksterhuis, A., … van't Veer, A. (2013). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217–224, http://dx.doi.org/10.1016/j.jesp.2013.10.005.
- Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6, 635–640. https://doi.org/10.3758/BF03212972.
- Calogero, R. M. (2013). Objects don't object: Evidence that self-objectification disrupts women's social activism. *Psychological Science*, *24*(3), 312–318.

- Camerer, C.F., Dreber, A., Holzmeister, F., … Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour,* 2**,** 637–644, https://doi.org/10.1038/s41562-018-0399-z.
- Caron, E. E., Reynolds, M. G., Ralph, B. C. W., … Smilek, D. (2020). Does posture influence the Stroop effect? *Psychological Science*, *31*(11), 1452–1460, https://doi.org/10.1177/0956797620953842.
- Chambers, C. (2019). What's next for registered reports? *Nature*, 573, 187-189.
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610, https://doi.org/10.1016/j.cortex.2012.12.016.
- Chambers, C. D., & Tzavella, L. (2021, June 28). The past, present, and future of Registered Reports. https://doi.org/10.31222/osf.io/43298.
- Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *PNAS*, 106(21), 8719-8724, https://doi.org/10.1073/pnas.0900234106.
- Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: Three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman's (2013) Findings. *Psychological Science*, *32*(4), 611–621, https://doi.org/10.1177/0956797620972367.
- De Wilde, M., Casini, A., Bernard, P., … Demoulin, S. (2020). Two preregistered direct replications of "objects don't object: evidence that self-objectification disrupts women's social activism". *Psychological Science*, *31*(2), 214–223.
- Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946, https://doi.org/10.1177/0956797612464058.

- Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology*, *5*(1): 46.
- Forster, S., & Lavie, N. (2009). Harnessing the wandering mind: The role of perceptual load. *Cognition*, 111, 345-355, https://doi.org/10.1016/j.cognition.2009.02.006.
- Frank, M.C., Bergelson, E., Bergmann, C., … Yurovsky, D. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy*, 22, 421-435, https://doi.org/10.1111/infa.12182.
- Fredrickson, B. L., & Roberts, T. A. (1997). Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly*, 21, 173–206.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103(6), 933-948, https://doi.org/10.1037/a0029709.
- Gervais, W. M., McKee, S. E., & Malik, S. (2020). Do religious primes increase risk taking? Evidence against "anticipating divine protection" in two preregistered direct replications of Kupor, Laurin, and Levav (2015). *Psychological Science*, *31*(7), 858–864, https://doi.org/10.1177/0956797620922477.
- 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.
- Halfmann, E., Bredehöft, J., & Häusser, J. A. (2020). Replicating roaches: A preregistered direct replication of Zajonc, Heingartner, and Herman's (1969). Social-Facilitation Study. *Psychological Science*, *31*(3), 332–337, https://doi.org/10.1177/0956797620902101.
- Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, 2, 793-796, https://doi.org/10.1038/s41562-018-0444-y.

- Hardwicke, T. E., & Wagenmakers, E. (2021, April 23). Preregistration: A pragmatic tool to reduce bias and calibrate confidence in scientific research. https://doi.org/10.31222/osf.io/d7bcu.
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, 5:180448, https://doi.org/10.1098/rsos.180448.
- Hardwicke, T. E., Thibault, R. T., Kosie, J. E., … Ioannidis, J. P. A. (2021). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, 1-13.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124, https://doi.org/10.1371/journal.pmed.0020124.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532, https://doi.org/10.1177/0956797611430953.
- Jost, J. T., Banaji, M., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and nonconscious bolstering of the status quo. *Political Psychology*, 25, 881–919.
- Kidwell, M. C., Lazzarević, L. B., Baranski, E., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, 14(5): e1002456, https://doi.org/10.1371/journal.pbio.1002456.
- Kupor, D. M., Laurin, K., & Levav, J. (2015). Anticipating divine protection? Reminders of God can increase nonmoral risk taking. *Psychological Science*, *26*(4), 374–384, https://doi.org/10.1177/0956797614563108.

- Lebel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, 113(2), 254-261, http://dx.doi.org/10.1037/pspi0000106.
- Levinson, D. B., Smallwood, J., & Davidson, R. J. (2012). The persistence of thought: Evidence for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science*, 23(4), 375-380, https://doi.org/10.1177/0956797611431465.
- Machery, E. (2020). What is replication?. *Philosophy of Science*, 87(4), 545-567, https://doi.org/10.1086/709701.
- Meier, M. E. (2019). Is there a positive association between working memory capacity and mind wandering in a low-demand breathing task? A preregistered replication of a study by Levinson, Smallwood, and Davidson (2012). *Psychological Science*, 30(5), 789-797, https://doi.org/10.1177/0956797619837942.
- Miller, S. L., & Maner, J. K. (2011). Sick body, vigilant mind: the biological immune system activates the behavioral immune system. *Psychological Science*, *22*(12), 1467–1471, https://doi.org/10.1177/0956797611420166.
- Montoya, A. K., Krenzer, W. L. D., & Fossum, J. L. (2021). Opening the door to registered reports: Census of journals publishing registered reports (2013–2020). *Collabra: Psychology*, 7(1): 24404, https://doi.org/10.1525/collabra.24404.
- Munafò, M. R. (2017). Improving the efficiency of grant and journal peer review: registered reports funding. *Nicotine & Tobacco Research*, 19(7), 773.
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511-534, https://doi.org/10.1146/annurev-psych-122216-011836.
- Newman, R. S., & Hussain, I. (2006). Changes in preference for infant-directed speech in low and moderate noise by 4.5- to 13-month-olds. *Infancy,* 10, 61–76. https://doi.org/10.1207/s15327078in1001_4.

- Nock, M. K., Park, J. M., Finn, C. T., … Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511–517, https://doi.org/10.1177/0956797610364762.
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, 1(3), 1-8, https://doi.org/10.1371/journal.pbio.3000691.
- Nosek, B. A., Beck, E. D., Campbell, L., … Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815-818, https://doi.org/10.1016/j.tics.2019.07.009.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor D. T. (2017). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2021, February 9). Replicability, robustness, and reproducibility in psychological science. https://doi.org/10.31234/osf.io/ksfvq.
- Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 229-237, https://doi.org/10.1177/2515245920918872.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-1 – aac4716-8.
- Ozer, D. J., & Benet-Martìnez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401-421, https://doi.org/10.1146/annurev.psych.57.102904.190127.
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

- Pham, M. C., & Oh, T. T. (2021). Preregistration is neither suffcient nor necessary for good science. *Journal of Consumer Psychology*, 31(1), 163-176.
- Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drugs Discovery*, 10, 712, https://doi.org/10.1038/nrd3439-c1.
- Reich, J., Gehlbach, H., & Albers, C. J. (2020). "Like upgrading from a typewriter to a computer": Registered reports in education research. *AERA Open*, 6(2), 1-6, https://doi.org/10.1177/2332858420917640.
- Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE*, 7(3), 1-5, https://doi.org/10.1371/journal.pone.0033423.
- Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass,* 14: e12633, https://doi.org/10.1111/phc3.12633.
- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178.
- Rosenbaum, D., Mama, Y., & Algom, D. (2017). Stand by your Stroop: Standing up enhances selective attention and cognitive control. *Psychological Science*, *28*(12), 1864–1867, https://doi.org/10.1177/0956797617721270.
- Schaller, M., Miller, G. E., Gervais, W. M., Yager, S., & Chen, E. (2010). Mere visual perception of other people's disease symptoms facilitates a more aggressive immune response. *Psychological Science*, 21, 649–652.
- Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, 4(2), 1-12,https://doi.org/10.1177/25152459211007467.

- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47, 609-612, https://doi.org/10.1016/j.jrp.2013.05.009.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366, https://doi.org/10.1177/0956797611417632.
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569, https://doi.org/10.1177/0956797614567341.
- Smallwood, J. & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, 103(6), 946-958, https://doi.org/10.1037/0033-2909.132.6.946.
- Soderberg, C. K., Errington, T. M., Schiavone, S. R., … Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, 5, 990-997, https://doi.org/10.1038/s41562-021-01142-4.
- Soto, C. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, 30(5), 711-727, https://doi.org/10.1177/0956797619831612.
- Stawarczyk, D., Majerus, S., Maquet, P., & D'Argembeau, A. (2011). Neural correlates of ongoing conscious experience: Both task-unrelatedness and stimulus-independence are related to default network activity. *PLoS ONE*, 6(2): e16997, https://doi.org/10.1371/journal.pone.0016997.
- Szollosi, A., Kellen, D., Navarro, D. J., … Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, 24(2), 94-95.
- Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: Direct replication of the predictive validity of the suicide–implicit association test. *Psychological Science*, 31(1), 65–74, https://doi.org/10.1177/0956797619893062.

- The ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24-52, https://doi.org/10.1177/2515245919900809.
- Toth, A. A., Banks, G. C., Mellor, D., … Borns, J. (2020). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, 36, 553-571, https://doi.org/10.1007/s10869-020-09695-3.
- Tybur, J. M., Jones, B. C., DeBruine, L. M., Ackerman, J. M., & Fasolt, V. (2020). Preregistered direct replication of "sick body, vigilant mind: the biological immune system activates the behavioral immune system". *Psychological Science*, 31(11), 1461–1469, https://doi.org/10.1177/0956797620955209.
- van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology,* 67, 2–12, http://dx.doi.org/10.1016/j.jesp.2016.03.004.
- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity and progress. *Perspectives on Psychological Science*, 13(4), 411-417, https://doi.org/10.1177/17456916177518.
- Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475, https://doi.org/10.1037/a0036731.
- Wagenmakers, E., Wetzels, R., & Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426-432, https://psycnet.apa.org/doi/10.1037/a0022790.
- Wagenmakers, E., Wetzels, R., Borsboom, D., Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632-638, https://doi.org/10.1177/1745691612463078.

- Walmsley, J., & O'Madagain, C. (2020). The worst-motive fallacy: A negativity bias in motive attribution. *Psychological Science*, *31*(11), 1430–1438.
- Yamada, Y. (2018) How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology*, 9: 1831.
- Zajonc, R. B., Heingartner, A., & Herman, E. M. (1969). Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology, 13*(2), 83–92, https://doi.org/10.1037/h0028063.
- Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication streaming. *Behavioral and Brain Sciences*, 41, 1-61.

Sitografia

- https://aspredicted.org/
- https://github.com/crsh/prereg
- https://help.osf.io/hc/en-us/articles/360019738834-Create-a-Preregistration#Submit-your-preregistration
- https://manybabies.github.io/
- https://osf.io/2mh48/
- https://osf.io/43qtn
- https://osf.io/64ct2/
- https://osf.io/6w8qt/
- https://osf.io/7d3xh/
- https://osf.io/8cwgx/
- https://osf.io/93znh/
- https://osf.io/9j6d7/
- https://osf.io/c7t6k

- https://osf.io/dg9m4
- https://osf.io/gf7vh
- https://osf.io/jea94/
- https://osf.io/k2dbf/
- https://osf.io/m28xv
- https://osf.io/mjrpy/
- https://osf.io/nx2sv/
- https://osf.io/pfdyw/
- https://osf.io/rkfq5/
- https://osf.io/tvyxz/wiki/home/
- https://osf.io/wcy2p/
- https://osf.io/x5w7h/wiki/home/
- https://osf.io/zep2b/
- https://prereg-psych.org/index.php/rrp
- https://www.apa.org/science/about/psa/2011/12/diederik-stapel
- https://www.cos.io/initiatives/prereg
- https://www.cos.io/initiatives/registered-reports
- https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/food-and-drug-administration-amendments-act-fdaaa-2007
- https://www.nationalacademies.org/news/2019/05/new-report-examines-reproducibility-and-replicability-in-science-recommends-ways-to-improve-transparency-and-rigor-in-research#:~:text=Reproducibility%20means%20obtaining%20consistent%20computational,has%20obtained%20its%20own%20data
- https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=rrsp20

- https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Appendice A

Modulo di preregistrazione *AsPredicted*.

AsPredicted registration:

1. Have any data been collected for this study already? (optional)
   - Yes, at least some data have been collected for this study already
   - No, no data have been collected for this study yet

2. What's the main question being asked or hypothesis being tested in this study? (optional)

3. Describe the key dependent variable(s) specifying how they will be measured. (optional)

4. How many and which conditions will participants be assigned to? (optional)

5. Specify exactly which analyses you will conduct to examine the main question/hypothesis. (optional)

6. Any secondary analyses? (optional)

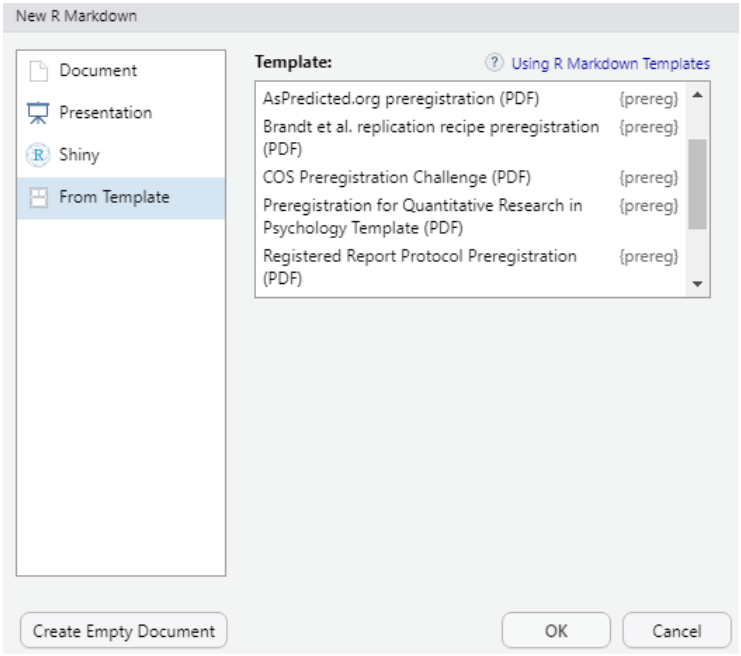
7. How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined. (optional)

8. Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?) (optional)

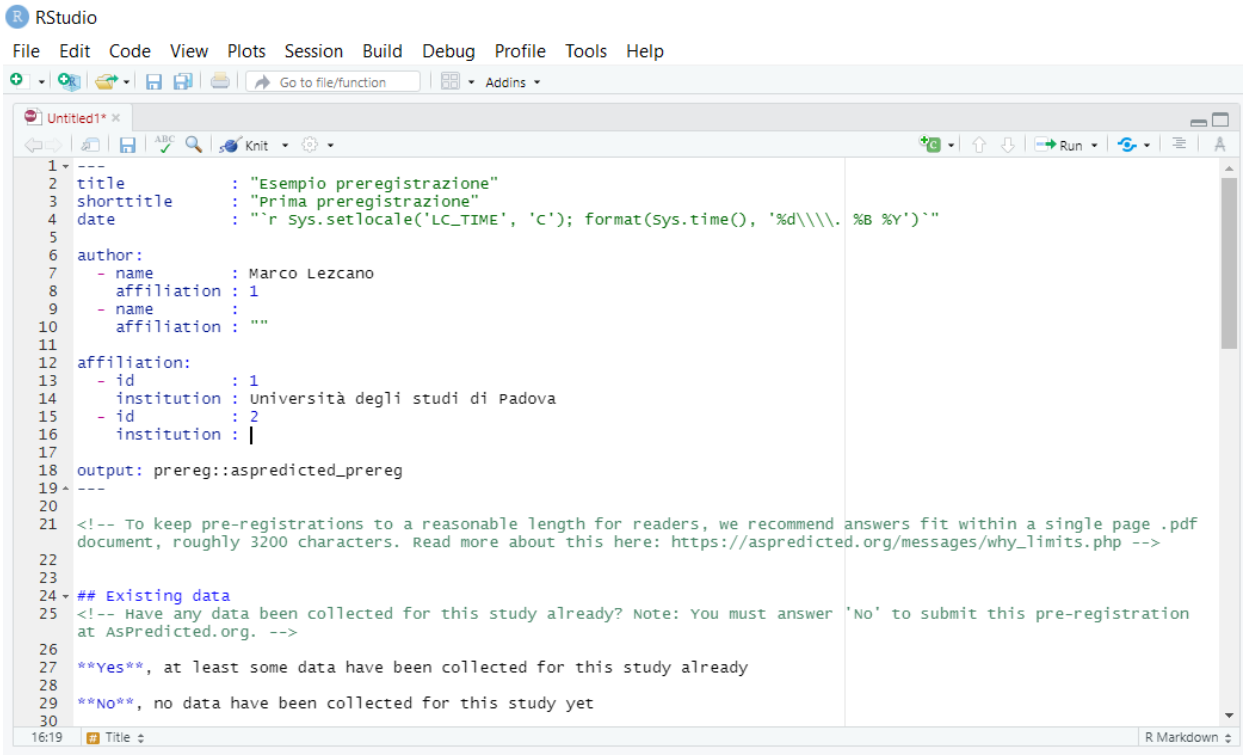
Appendice B

Come utilizzare RStudio per preregistrare uno studio (guida per windows).

1. Installare nel proprio pc: una versione di R non precedente alla 2.11.1, una versione di RStudio non precedente alla 0.99.441 ed il programma MiKTeX;
2. Aprire RStudio ed installare i pacchetti rmarkdown e prereg;
3. Cliccare su File > New File > Rmarkdown > From Template e scegliere il modello di preregistrazione che si vuole utilizzare;



4. Dopo che si è selezionato il modello di preregistrazione compilarlo ed alla fine della compilazione premere sul pulsante Knit;



5. Selezionare la cartella dove salvare il file PDF contenente la preregistrazione;
6. Verranno generati una serie di file, tra cui il PDF contenente la preregistrazione che è stata compilata in RStudio;

Preregistration

Esempio preregistrazione

Marco Lezcano[1],

[1] Università degli studi di Padova

[2]

20. September 2021

Existing data	**Yes**, at least some data have been collected for this study already
	No, no data have been collected for this study yet

Hypothesis

Example: A month-long academic summer program for disadvantaged kids will reduce the drop in academic performance that occurs during the summer.

Appendice C

Preregistrazione di Crawford, J. T., & Ruscio, J. (2021). Asking people to explain complex policies does not increase political moderation: Three preregistered failures to closely replicate Fernbach, Rogers, Fox, and Sloman's (2013) findings. *Psychological Science*, *32* (4), 611–621.

Preregistration Template from AsPredicted.org

Have any data been collected for this study already?

No, no data have been collected for this study yet

What's the main question being asked or hypothesis being tested in this study?

This is a preregistration of Studies 2 and 3 from Fernbach et al. (2013). In Study 2, the original study showed that asking people to create mechanistic explanations for their political opinions (compared to simply providing reasons for their political opinions) led to lower estimates of their understanding of the political issue itself, and more moderated stances on the issue. Using a similar procedure, Study 3 showed that mechanistic explanations (compared to reasons) led to less donations toward a politically sympathetic organization. We seek to replicate these findings in two separate samples (one replicating Study 2, the other replicating Study 3). We are also including an additional set of measures in the replication of Study 3, to be included following the original measures. Specifically, we are testing the hypothesis, derived from Fernbach et al. (2013), that mechanistic explanations (relative to reasons) will lead to a decrease in negative attitudes toward ideological outgroups generally, and relative to ideological ingroups specifically.

Describe the key dependent variable(s) specifying how they will be measured.

Study 2: The dependent variables are changes in understanding and changes in position extremity. Following Fernbach et al., they are pre- and post-measures of each, which will be submitted to repeated measures ANOVAs. Study 3: The original dependent variable in Fernbach et al. was a binary measure of whether the participants decided to donate to an ideological ingroup organization or not. The additional dependent measures are feeling thermometer ratings and social distance ratings of each of the four target groups from the study, along with liberals and conservatives (so, two separate items for each of the six targets groups). Following other work (e.g., Crawford et al., 2017), the feeling thermometer and social distance ratings will be standardized and combined to form a single prejudice measure toward each target (presuming they are adequately positively correlated, at least above a small effect size). We will examine whether there is an influence of the independent variable on each item individually, and we will also create a difference score item, reflecting the bias in favor of a group over its opposite (i.e.., supporters vs. components of a flat tax; supporters vs. opponents of cap and trade; the liberal and conservative targets are included for exploratory purposes).

How many and which conditions will participants be assigned to?

Study 2: Participants are assigned to either the mechanistic or reasons condition. Within those

conditions, they are then randomly assigned to one of three possible pairs of issues. This follows Fernbach et al.'s original procedures. Study 3: Participants are assigned to provide either a mechanistic or reasons explanation for one of two issues. This follows Fernbach et al.'s original procedures.

Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Study 2: Following Fernbach et al., we will conduct repeated measures ANOVAs on understanding and political extremity, and a correlation between changes in understanding and extremity. (These were performed as replications of Study 1). Although Fernbach et al do not specify, it appears they conducted mixed ANOVAs, with condition as a between subjects variable and changes in understanding and extremity as within subject variables. We will conduct these analyses as well. Fernbach et al also note that they performed exploratory analyses by content coding responses in the reasons condition after observing an unexpected effect on understanding. If we observe the same effect, we will perform the coding as outlined in their supplemental materials. Study 3: Following Fernbach et al., we will conduct a logistic regression on the binary donation variable, with condition, initial extremity of policy support, and the interaction as independent variables. For the intergroup attitudes measures, we will conduct linear regression models on the standardized and combined prejudice measures. Any significant interactions will be followed up to determine the nature of the interaction through spotlight tests at the highest and lowest levels of extremity.

Any secondary analyses?

NA

How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

Following Simonsohn's (2015) recommendation, we aim to collect at least 2.5 times the original sample size. This translates to at least 353 participants for Study 2, and at least 253 participants for Study 3. We may oversample from these minimum sample sizes.

Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

Following Fernbach et al., we will exclude participants who fail the attention check in both studies. We are using a similar attention check to that used in the original studies, although the original authors could not provide the exact item. Regarding exploratory analyses, in Study 3, we are collecting data on how liberal or conservative the two issues are perceived. We will test the hypothesis that the original effects are most pronounced among participants who accurately perceive the ideological meaning of the issue tested by including perceived ideological meaning as an independent variable in the model, along with its interaction terms. To the best of our ability, we have followed the procedure and used the materials from the original studies. We obtained as many of the original materials as possible from the original authors. The only exceptions that we are aware of to the original protocols are for Study 3. Specifically, a) we are adding the intergroup attitude measures. However, these are included after all key study items (though before the demographics), meaning they cannot have influence over the other measures; and b) Fernbach et al. had actually identified organizations for which to donate money in Study 3, but could not recall the names of these organizations. Instead, at the end of the study, we will tell participants that they will each personally receive the bonus payment of 20 cents, regardless of what they chose to do with the money. This change to the protocol should not create a psychologically different experience between the original and replication, as participants in both would anticipate that the researchers

would follow their request for how to distribute the money.

Preregistrazione di Gervais, W. M., McKee, S. E., & Malik, S. (2020). Do religious primes increase risk taking? Evidence against "anticipating divine protection" in two preregistered direct replications of Kupor, Laurin, and Levav (2015). *Psychological Science*, *31* (7), 858-864.

The Nature of the Effect

Verbal description of the effect I am trying to replicate

We seek to replicate the finding that priming God (using religious words) leads to participants taking more non-moral risks than the control condition (who were primed with non-religious words).

It is important to replicate this effect because

It is important to replicate this effect because priming techniques like those used in this set of studies have come under some heavy questioning in recent years. Also, this paper represents repeated apparently successful applications of the primes and were published in a top journal. Finally, most reported effects in this paper are just barely significant, raising the possibility of nonrobustness and/or biased reporting. A replication can ensure the robustness of these findings.

The effect size of the effect I am trying to replicate is

For Study 1a – Cohen's D = .574; For Study 1b – Cohen's D = .323

The confidence interval of the original effect is

For Study 1a – Cohen's D = .574 [.04, 1.09]; For Study 1b – Cohen's D = .323 [.04, .60]

The sample size of the original effect is

For Study 1a the original sample size was 61. For Study 1b the original sample was 202

Where was the original study conducted? (e.g., lab, in the field, online)

Online – Mturk

What country/region was the original study conducted in?

Online – so American sample?

What kind of sample did the original study use? (e.g., student, Mturk, representative)

Mturk

Was the original study conducted with paper-and-pencil surveys, on a computer, or something else?

On a computer

Designing the Replication Study

Are the original materials for the study available from the author?

yes

I know that assumptions (e.g., about the meaning of the stimuli) in the original study will also hold in my replication because

Mixed. The authors provided some stimuli but could not provide experimental scripts or details on delivery of instructions. We worked with them to ensure a high-fidelity replication. We are using the same materials and sampling technique used in the original; further the paper is quite recent. There is no a priori to expect any of the (same) materials to hold a new meaning in a sample drawn from the same population, a few years later.

Location of the experimenter during data collection

Not present during data collection, as it will be completed online.

Experimenter knowledge of participant experimental condition

Study 1a = random assignment. Researchers do not know in advance what condition participants are assigned to. Study 1b = random assignment. Researchers do not know in advance what condition participants are assigned to.

Experimenter knowledge of overall hypotheses

For both Study 1a and 1b, the experimenter (PI) knows the hypotheses, however, as this is study is online, they cannot influence participants responses.

My target sample size is

Study 1a = 600; Study 1b = 600

The rationale for my sample size is

In both study 1a and 1b, we wish to exceed the "rule of thumb" of collecting a sample that is at least 2.5x as large as the original study sample size. Beyond that, we sought to maximize our sample size, given our available resources.

Documenting Differences between the Original and Replication Study

The similarities/differences in the instructions are

Close

The similarities/differences in the measures are

Close

The similarities/differences in the stimuli are

Exact

The similarities/differences in the procedure are

Close

The similarities/differences in the location (e.g., lab vs. online; alone vs. in groups) are

Exact

The similarities/difference in remuneration are

No response

The similarities/differences between participant populations are

Exact

What differences between the original study and your study might be expected to influence the size and/or direction of the effect?

The larger sample size in our study will yield more precise estimates, but should not affect size or direction.

I have taken the following steps to test whether the differences listed in the previous question will influence the outcome of my replication attempt

N/A

Analysis and Replication Evaluation

My exclusion criteria are (e.g., handling outliers, removing participants from analysis)

In both Study 1a and 1b, we will remove participants who fail an attention check embedded in each of the studies. Additionally, we will exclude participants in Study 1a and 1b who failed to properly complete the sentence scramble task (the IV), by either using all 5 words, or creating sentences which are grammatically incorrect. In addition, we will omit participants who do not use the target prime words in crucial trials in the experimental condition.

My analysis plan is (justify differences from the original)

We will use a t-test to determine if there is a difference in risk taking between conditions in both studies (in line with the original study). We also want to check to see if the effects are moderated by a particiants belief in God. This analysis was run by the original research team in later studies in the same paper, so we wish to check if the effect holds here as well. We wish to examine the data on financial risk in an exploratory manner. Although the original research team did not look at this data, we think "risk responses" for the financial questions will be significantly lower than other domains. Finally, we will use "small telescopes" and Bayesian analyses to examine whether our replication is meaningfully different from the originals.

A successful replication is defined as

i. Producing directionally same significant effects as the original,

ii. Further support will come from the "small telescopes" approach

iii. We will also present a Bayes Factor analyses comparing a point null hypothesis to an alternative hypothesis based on the effect sizes obtained in the original study

iv. Finally, if we can obtain original raw data, we will perform a Bayesian hierarchical model examining the experimental effects nested within sample. This will provide a full posterior distribution of the pooled effect size, as well as differences between replication and original samples. Failing this, we will present a meta-analysis pooling original and replication efforts.

Preregistrazione di Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511–517.

Study Information

Title

Does Posture Influence the Stroop Effect? – modified preregistration

Authors

Description

Rosenbaum, Mama, and Algom (2017) recently reported the intriguing finding that the Stroop effect is smaller when participants complete the task while standing than while sitting. These findings are important for several reasons: First, it is often difficult to find conditions that lead to a substantial reduction in the magnitude of the Stroop effect, and this has led to the prevailing view that reading the to-be-ignored words is automatic. In this theoretical context, the finding that the Stroop effect can be reduced simply by having participants stand has important theoretical implications as it challenges the prevailing view of the automaticity of the effect. Second, standing workstations are now rapidly gaining in popularity (Cooley & Pedersen, 2013) and Rosenbaum et al.'s findings provide important data showing that standing while working might induce an additional cognitive load. We found Rosenbaum et al.'s (2017) findings compelling, yet not completely consistent with findings from several prior studies in the literature. Specifically, previous examinations of the influence of standing vs. sitting on the Stroop task have drawn the conclusion that standing has no substantial effect on Stroop performance (e.g., Bantoft, et al., 2016). On the other hand, in a recent study published in May of this year, Smith, Davoli, Knapp and Abrams (2019) were able to replicate Rosenbaum et al.'s findings. Importantly, in the available studies, the Stroop task was not implemented in the same way as it was in Rosenbaum et al. In addition, we noted that the Rosenbaum et al (2017) did not include a neutral condition, which makes it difficult to know whether changes in the Stroop effect were due to changes in facilitation on congruent trials, or changes in interference on incongruent trials. Because of these issues, we thought it important to undertake a replication and an extension of the original findings with a larger sample of participants (N= 122, more than doubling Rosenbaum et al.'s large sample size). In all, we conducted four extended replications of Rosenbaum et al.'s (2017) study to address these issues. We first began our replication attempts prior to the publication of Rosenbaum's et al. (2018) corrigendum paper. Critically, across four attempted replications [with relatively large sample sizes – between N = 78 to N = 108] of Rosenbaum et al.'s study we were unsuccessful in finding an effect of posture (sit vs. stand) on the magnitude of the Stroop effect. Admittedly, all four of our experiments differ slightly from Rosenbaum's methodological design in one more of the following ways: In one experiment we included neutral trials (Experiment 1); in all four experiments we increased the number of trials; in several experiments we replaced Rosenbaum et al.'s word stimuli "BROWN" with the word stimuli "YELLOW" (Experiment 3 and 4); in two experiments we relied on manual responses (Experiment 3 and 4); and in one experiment we had people standing on one foot in the standing condition

(Experiment 4). Here, we propose a well-powered additional replication that strictly follows Rosenbaum et al. (2017) methodology.

Hypotheses

Our aim is to closely replicate Rosenbaum et al.'s (2017) studies to determine the impact of Posture on the magnitude of the Stroop effect. Our main hypotheses are that 1) participants will perform better on congruent trials than on incongruent trials, and more critically 2) based on our prior failed replications of Rosenbaum et al., we expect that posture will not influence the magnitude of the Stroop Effect.

Design Plan

Study type

Experiment – A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This 98arie s known as an intervention experiment and includes randomized controlled trials.
Blinding

- No blinding is involved in this study.

Is there any additional blinding in this study?

The experimenter will be ignorant of the goals and hypothesis of the study.

Study design

Both Posture (sitting vs. standing) and Congruency (congruent vs. incongruent trials) are within-participant factors. Participants complete one block of trials while sitting and one block of trials while standing. The order of the blocks is counterbalanced across participants. Each block contains a random assortment of congruent and incongruent trials.

No files selected

Randomization

Participants will be randomly assigned to one of two counterbalances (sitting first vs. standing first). Congruent and incongruent trials will be randomized.

Sampling Plan

Existing Data

Registration prior to creation of data
Explanation of existing data

No response

Data collection procedures

Participants. We will be recruiting undergraduate students from the University of Waterloo. The study should take no longer than 30 minutes and participants will be awarded 0.5 credit for their participation. There are no exclusion criteria. Stimuli. We will use the colour words described in Rosenbaum's et al.'s original paper: "RED", "GREEN", "BLUE", and "BROWN". The congruent stimuli

will be presented as follows: the words “RED” presented in the colour red (RGB 150, 0, 0), “GREEN” presented in the colour green (RGB: 27, 111, 27), “BLUE” presented in the colour blue (RGB: 0, 0, 150), and “BROWN” presented in the colour brown (RGB: 68, 47, 41). The Incongruent stimuli will include all other combinations of the colour words and hues. The stimuli will be displayed in uppercase Miriam font, on a light grey (RGB: 122, 122, 122) background. The viewing distance in both the sitting and standing conditions will be approximately 60 cm and the words will subtend 0.57 degrees of visual angle in height (as reported by Rosenbaum et al.); the width will vary based on word length but will be constrained by the predetermined height (1.3°) and the font type (Miriam). Participants will be presented with 72- colour-word Stroop stimuli, half (36) of which will be congruent and half (36) of which will be incongruent (with nine repetitions of each hue in a given condition). Participants will be randomly assigned to the testing conditions (starting sitting vs. starting standing) and this order will be counterbalanced. Participants will be brought into the experimental room. All participants will be presented the same instructions on the screen and the experimenter (who is ignorant of the goals and hypothesis of the study) will read the instructions out loud to the participant. As mentioned above, participants will complete one experimental condition sitting and the other standing and the order of posture condition will be counterbalanced across participants. On each trial participants will be asked to respond “as quickly and accurately as possible” to the hue while ignoring the meaning of the letter string. On each trial, the letter string will appear and remain on the screen until a response is made, after which 99arie s be replaced by a grey screen for 1000 ms. Apparatus: The experiment will be programmed using E-prime 3 programming software and will be run on a desktop PC. Stimuli will be presented on a Dell 2007 WFP monitor with the display resolution set to 1680 x 1050 True colour (32 bit) at 59Hz. The computer and monitor will be placed on an Ikea BEKANT desk (https://www.ikea.com/ca/en/catalog/products/S29022520/), which can be electrically adjusted in height to accommodate the height of the participants in both conditions so that the position of the arms and head follow the accepted ergonomic guidelines (Canadian Centre for Occupational Health and Safety, 2019). In the sitting condition, participants will sit on a fixed plastic chair with a metal base. The computer monitor will be adjusted so that the center of the screen will be at eye level for participants. Vocal response times will be collected using a microphone and voice key. Vocal responses will be recorded for a later coding of response accuracy.

No files selected

Sample size

The study will consist of a Bayesian sequential design and involve testing up until the Bayes factor is 5 for the critical Posture by Congruency interaction.

Sample size rationale

Our primary hypothesis involves testing an interaction across two experimental conditions (Posture and Congruency). A power calculation in G*Power with a criterion of .95 power based on the partial eta ($\eta p2$ = .155) obtained by Rosenbaum et al. (2017), suggests a sample size of 20 would be sufficient to detect the effect. However, as mentioned above, in the initial two replications conducted at the University of Waterloo, we used a much larger sample of 122 and were unable to obtain the effect. Based on the suggestions of the reviewers of our initial experiments, we have opted to conduct our experiment using the Bayesian sequential analysis method and to stop the sampling process once a Bayes factor of 5 is reached for the critical Posture by Congruency interaction.

Stopping rule

We will use the Bayesian sequential analysis method and stop the sampling process once a Bayes factor of 5 is reached for the critical Posture by Congruency interaction.

Variables

Manipulated variables

No response

No files selected

Measured variables

The Response Time and Accuracy of the vocal responses to the Stroop stimuli will be measured as the main outcome variables.

No files selected

Indices

No response

No files selected

Analysis Plan

Statistical models

Our primary analysis will consist of a mixed measures ANOVA with Posture (Sitting vs. Standing) and Congruency (Incongruent vs. Congruent) as within-participant factors and Counterbalance (Sitting first vs. Standing First) as a between-participant factor. We will conduct the three-way ANOVA and focus on the Posture by Congruency interaction as the key test of the hypothesis. Should there be a significant 3-way interaction, we will describe it and analyze each order using follow-up ANOVAs just for the sake of completeness. If we fail to find a significant Posture by Congruency interaction, we will consider the effect Rosenbaum et al. Reported to not be replicated. We will also use Bayesian analyses to test the expected null Posture by Congruency interaction.

No files selected

Transformations

No response

Inference criteria

For the ANOVA, we will use a p-value of .05 as our standard significance cut-off. For the Bayesian analyses, we will collect data to obtain a Bayes factor of 5.

Data exclusion

We will exclude participants if the levels of missing data (due to a failure to record vocal responses) within each individual participant's file is higher than 20%, since the resulting data from these participants would be based on too few trials to generate a stable mean estimate of reaction time. We will excluded trial data that is unusable due to: o Hardware failures o Premature triggering of the

voice key resulting from response artifacts (e.g., coughing, sneezing, breathing, aberrant vocal response). When analyzing the RT data, incorrect answers will also be excluded. RT data will also be submitted to the widely used recursive data trimming procedure (Van Selst & Jolicoeur, 1994), whereby the data for each participant and each condition will be trimmed separately so that a disproportional amount of data is not removed from a given condition.

Missing data

We will exclude participants if the levels of missing data (due to a failure to record vocal responses) within each individual participant's file is higher than 20%. Participants missing less than 20% of the overall data will remain within the analysis.

Exploratory analysis

No response

Other

Other

No response

Preregistrazione di Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: Direct replication of the predictive validity of the suicide–implicit association Test. *Psychological Science*, *31*(1), 65–74.

Project working title: Does the Implicit Association Test Prospectively Predict Suicidal Behavior? Direct Replication of Nock, Park, Finn, Deliberto, Dour, & Banaji (2010)

Authors: Nina Tello[1, 2], Gina Harika-Germaneau[2], Wilfired Serra[2], Armand Chatard[1, 2], & Nematollah Jaafari[2]

Affiliation: [1] Poitiers University, CNRS, France. [2] Hospital Center Henri Laborit, Poitiers, France.

C. Hypotheses Description of essential elements

1. Describe the (numbered) hypotheses in terms of directional relationships between your (manipulated or measured) variables.

In our first hypothesis (Hypothesis 1), we predict, as Nock et al. (2010), that patients who present themselves to the emergency department after a suicide attempt would have a significantly stronger implicit association between self and death than control patients. In a second hypothesis (Hypothesis 2), we will test if the strength of this association between self and death would predict future suicide attempt over 6 months beyond other clinical predictors (depressive disorder, multiple suicide attempts, suicide ideation, clinician and patients prediction).

2. For original research, add rationales or theoretical frameworks for why a certain hypothesis is tested.

Our study is a direct replication of Nock et al. (2010). Since, most previous studies relied on subclinical samples, our study aims to overcome this shortcoming by closely and independently replicating Nock et al. 's (2010) original research with a clinical population. Independent replications are important because replications are more likely successful when leaded by the same team (Makel, Plucker, & Hegarty, 2012).

B. Methods

Description of essential elements Design

C. Independent variables with all their levels

• Suicide attempt status

Patients with no recent suicide attempt (controls) versus patients with a recent suicide attempt (in the week before the emergency)

C. Dependent variables, or variables in a correlational design

- Performance on the IAT
- Presence of a suicide attempt in 6 months follow up period

C. Third variables acting as covariates or moderators

- any depressive disorder (covariate)

- multiple suicide attempts (covariate)
- suicide ideation (covariate)
- clinician prediction (covariate)
- patient prediction (covariate)

Planned sample

4. If applicable, describe pre-selection rules.

We will use the same inclusion an exclusion criteria than Nock et al. (2010). We will include any patients that will be at least 18 years old, and have an absence of impairment that could affect the ability to comprehend and participate to the study (inability to speak French, cognitive impairment, agitated or violent behaviors).

For Hypothesis 2, as in Nock et al.'s study (2010), only patients with a life history of previous suicide attempts will be selected in the analysis.

5. Indicate where, from whom and how the data will be collected.

We will recruit 162 participants from the psychiatric emergency of Poitiers Hospital. A member of the psychiatric clinical staff will first evaluate patients and if necessary they will be hospitalized. Then we will describe them the study and propose them to participate. Consenting patients will complete all measures in a small office in the hospital or in their hospital bed.

6. Justify planned sample size (if applicable, you can upload a file related to your power analysis here (e.g., a protocol of power analyses from G*Power, a script, a screenshot, etc.).
   The sample was determined in advance to have at least 80% power to replicate Nock, et al.'s (2010) original findings (https://osf.io/r3pfs/). The original study included 157 patients. The results showed that the 43 patients who had attempted suicide in the week prior to their hospitalization had higher suicide-IAT scores than the other 114 patients (control group) $t(157) = 2.46$, $p < .05$, Cohen's $d = 0.44$. This is the first key result we will seek to replicate in our study. Nock et al. (2010) also found that, after control for clinical predictors (any depressive disorder, multiple suicide attempts, suicide ideation, clinician prediction, and patient prediction) patients who had high suicide-IAT scores (continuous variable) were more likely to attempt suicide in the 6 months follow-up (OR = 30.68, 95CI [1.18 – 795.12]), compared to patients with low suicide-IAT scores. This is the second key result we will seek

to replicate. The observed effect sizes were medium to large in the original study[1] (Cohen's *d* = 0.44 and 1.88 for the first and the second key results respectively). We based our power analyses on the first key result reported by Nock et al. (2010) because it had the lowest effect size, thus a sample size sufficient to replicate this effect with 80% statistical power will necessary have sufficient power to replicate the second key result. Power analyses reveal that we will need a total sample size of 162 patients to replicate the first key results in a one-tailed T-test, with $\alpha = .05$, $1 - \beta = 0.80$, and a N1/N2 ratio = 2.65.

A one-sided T-test seems justified because 1) the direction of the prediction is unidirectional in a direct replication, 2) the prediction is clearly specified in advance in a preregistered study, and 3) we have no interest in and we will not try to interpret effects in the opposite direction.

Importantly, the group allocation ratio for the power analysis is selected based on the Nock's et al. (2010) study. There is no guarantee that this allocation ratio will be the same in this replication study. However, national suicide rates are higher in France than in the United States (World Health Organization, 2014). For that reason, we expect a greater proportion of suicidal patients compared to non-suicidal patients in this replication study. Thus, the power analysis provides a conservative estimate of the required total sample size.

Describe data collection termination rule.

We will stop data collection when 162 patients will be included in the study.

Exclusion criteria

C. Describe anticipated specific data exclusion criteria.

We will exclude patients who show evidence of cognitive impairment (e.g., severe psychotic symptoms or somnolence due to medication) and violent or agitated participants. We will also exclude patients that show an impairment, which could affect the ability to comprehend and participate to the study (inability to speak French, cognitive impairment, agitated or violent behaviors)

Based on D600 formula (Greenwald, Nosek, & Banaji, 2003) we will exclude all participants, which response latencies were too short (< 0,3 seconds) or too long (> 10 seconds) on the IAT.

We will define participants as outlier if their score deviates more than 3 standard deviations from the mean.

[1] To compute effect sizes, we used Lipsey and Wilson's (2001) Practical Meta-Analysis Effect size Calculator, correcting for unequal sample size, and DeCoster's (2012) Converting Effect Sizes Calculator.

Procedure

C. Describe all manipulations, measures, materials and procedures including the order of presentation and the method of randomization and blinding (e.g., single or double blind), as in a published Methods section.

Our study is an exact replication of Nock et al. (2010) study: the procedure and stimuli are strictly identical (except that they are translated to French). We will use the material of the original study sent to us by Professor Nock. This study has the approval of the ethical committee.

As Nock et al. (2010) we will include a brief cognitive impairment measure by asking several true/false questions about the study at the end of the consent form (In this study you will be asked to complete an interview, a brief computer tasks and a questionnaire? In this study you will be asked to answer several questions by telephone six months from now?). Patients, who respond correctly, will be allowed to participate. First, they will complete a death/life Implicit Association Task (Nock et al., 2010), then the Self-Injurious Thought and
Behaviors Interview (SITBI, Nock, et al., 2007) and finally the Beck Scale for Suicide Ideation (Beck & Steer, 1991). Six months later, they will be contacted by phone and complete the SITBI once again (Nock et al., 2007). We will also examine their hospital medical record to determine if they had returned to the hospital due to a suicide attempt during the 6 months period. As in Nock's et al. (2010) suicide, a suicide attempt will be considered to have occurred during the 6 months follow-up 105arie sas one of these two methods shows evidence of an attempt. It seems interesting to corroborate participants' selfreport with their medical record to have an objective indicator of suicide attempt.

Death/Self Implicit Association Test. The Implicit Association Test (IAT, Greenwald et al. 1998) is a computerized test of automatic mental associations between two concepts. IAT is based on reaction times to categorize stimuli that appear in the middle of the screen into one of two categories, in our study "Death" (i.e. *die, deceased, funeral, lifeless*, and *suicide*) and "Life" (i.e. *alive, survive, live, thrive*, and *breathing*) and/or into one of two attribute categories, in our study "Me" (i.e. *I, myself, my, mine,* and *self*) or "Not Me" (i.e. *they, them, their, theirs*, and *other*). Category names appear in the top left, and in the top right corners of the screen. Participants will have to decide if the stimulus belongs to the category on the left by pressing the "E" key, or to the category on the right by pressing the "I" key. The IAT comprises 7 blocks. Blocks 1, 2, and 5 are practice blocks. Blocks 3 and 4 are

congruent blocks where congruent categories share the same response key. “Congruence” is defined by the hypothesis being tested. In our study, we define “Me” and “Death”, as well as “Not me” and “Life” as congruent category pairs. In contrast, “Me” and “Life”, as well as “Not me” and “Death” were defined as incongruent category pairs. Blocks 6 and 7 are incongruent blocks where incongruent categories share the same response key. In our study, the IAT will be administered in French. Following the logic of the IAT, positive IAT scores represent a strong implicit association between self and death. The IAT scores will be computed using the D600 improved algorithm (Greenwald, Nosek, & Banaji, 2003).

Demographic and psychiatric factors. We will assess known demographic and psychiatric risk factors for suicide attempts in order to test the incremental predictive validity of the IAT. We will take into account participants’ age, sex, and main psychiatric diagnosis (evaluated by the Mini International Neuropsychiatric Interview, Sheehan et al., 1998).

History of suicidal behavior. We will determine group status (patients with a current suicide attempt versus controls) at baseline. We will also measure past history of suicidal behavior (number of past suicide attempts) at baseline, as prior suicide attempts are a strong predictor of subsequent suicide attempts (Nock, Borges, Bromet, Cha, Kessler, & Lee, 2008). History of suicidal behavior will be assessed via the Self Injurious Thoughts and Behaviors Interview (SITBI, Nock et al., 2007), a structured interview assessing participants history of suicidal and self-injury behaviors. This 169-question interview allows distinguishing among five different behaviors: suicidal ideation, suicide gesture, suicide attempt, thoughts of nonsuicidal self-injury, and non-suicidal self-injury. Every part of this questionnaire is devoted to one of these five behaviors and begins with a screening question. If the patient did the behavior all the questions of this part are asked. The questionnaire assesses the age

of the appearance of the behavior and the frequency of it. On a 5-point Likert Scale (0: low/little, 4: very much/ severe) patients have to respond to questions about the intensity of the behavior or thoughts, the function of the behavior (emotion regulation, communication) and the presence of this behavior in his/her relations. Methods used by patients will be collected using an openend question. This interview also assesses the pain and the drug use when the behavior was done. The interview has good reliability and validity (Nock et al., 2007). We will administer the French version of this interview, which we translated using the translation-back translation method (A researcher first translate to French the questionnaire and another researcher translate back this version in English, this allows us to compare between the two English version and to correct what was misunderstood).

Suicide Ideation. Patients will also complete the Beck Scale for Suicide Ideation (Beck & Steer, 1991) to assess the severity of suicide ideation. This is a 19-items questionnaire in which participants have to choose a response among three options. 107arie s be administered in French and translated using the translation-back translation method.

Clinical and patient predictions. We will ask the patient's primary therapist to predict the risk of a new suicide attempt ("Based on your clinical judgment and all that you know about this patient, if untreated, what is the likelihood that this patient will make a suicide attempt in the next six months?" (0-10, with 0 being no likelihood and 10 being very high likelihood). We will also assess the patient's own risk estimation by asking him or her "On a scale of 0 to 4, what is the likelihood that you will make a suicide attempt in the future?" In order to compare the predictive ability of the IAT to therapists' and patients' risk estimates, which are routinely used in psychiatric departments.

Follow-up assessment. We will assess the presence of suicide attempt during the 6 months follow-up period using two methods: we will re-administered the SITBI (Nock et al., 2007) by phone and we will examine patient's hospital medical record to determine whether they made a new suicide attempt and returned to the hospital. As in Nock's et al. (2010) suicide, a suicide attempt will be considered to have occurred during the 6 months follow-up 107arie sas one of these two methods shows evidence of an attempt. It seems interesting to corroborate participants' self-report with their medical record to have an objective indicator of suicide attempt.

C. Analysis plan

Confirmatory analyses

Describe the analyses that will test each main prediction from the hypotheses section.

We will perform the same analysis than Nock et al. (2010). We will use a T-test, to test our first hypothesis, in which we want to compare performance on the IAT between patients who did and those who did not make a suicide attempt immediately before their entrance to the psychiatric

emergency.

To test the second hypothesis in which we test whether IAT could predict future suicide attempt beyond other predictors, we will use hierarchical logistic regression. We will enter any depressive disorder and multiple suicide attempts, in a first step, suicide ideation, clinician prediction, and patient prediction, in a second step, and IAT scores (continuous variable), in a third step[2]. In this analysis, we will focus on patients with a lifetime history of suicide attempt, to test if IAT predicts suicide over 6 months.

Answer the following final questions:

Has data collection begun for this project?

Yes, data collection is underway. This research project is part of Nina Tello PhD Thesis. Data collection started on June 1st 2016 and it was expected to last at least two years. At the time we started with data collection, we were not sure that we could recruit enough patients to conduct this project to the end. So, we decided to preregister the project at the mid-term of data collection (when we were certain to have enough patients to terminate data collection within the two-year of Nina's thesis).

If data collection has begun, have you looked at the data? ○ No

The (estimated) start and end dates for this project are:

This project started on the June 1st 2016 and will end on August 2018.

Any additional comments before I pre-register this project (optional): ○ No

Preregistrazione di Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32* (7), 1169–1178.

Study Information

[2] As in Nock et al. (2010) study, we will also check if IAT dichotomous scores (IAT scores > 0 vs. IAT scores <= 0) could prospectively predict suicide attempt during the 6-month follow-up beyond the effect of other predictors.

Hypotheses

H* (SCORE focal test): A significant positive interaction between headline veracity (true or false headline) and treatment (accuracy induction) predicting likelihood to share, such that the treatment condition increases sharing discernment. H1: Prompting people to think about accuracy decreases the likelihood that they will be willing to share false information about COVID-19 on social media. H2: Prompting people to think about accuracy increases the likelihood that they will be willing to share true information about COVID-19 on social media.

Design Plan

Study type

Experiment – A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This 109arie s known as an intervention experiment and includes randomized controlled trials.
Blinding

- For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Is there any additional blinding in this study?

Participants will be assigned to one condition in the first study (treatment and control; explained below), and will be randomly assigned to one of 2 conditions in the second study (accuracy; see below). Participants do not know in advance what condition they will be assigned to.
Study design

This is a replication of study 2 of the paper "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention" (2020) by Gordon Pennycook, Jonathan McPhetres, Yunhao Zhang and David Rand. Identical to their study (as noted in their pre-registration), we ask the following question: does prompting people to think about accuracy decrease the likelihood that they will be willing to share false information about COVID-19 on social media? We will be running 2 studies: study 1 is a partial replication of study 1 from the original paper, and study 2 is a full replication of their study 2. Our study 1 only involves one accuracy condition, in which participants will be asked whether they believe a set of 15 false and 15 real news (COVID-19 related) headlines are accurate. The accuracy condition was one of the two conditions from Study 1 from the original paper, results from which (specifically: average accuracy scores for true and false headlines about COVID-19) were used for further analyses in their Study 2 (which we are replicating). In our study 2, participants will be randomly assigned to one of two conditions: 1) a control condition, where they indicate whether they would share a set of 15 false and 15 real news (COVID-19 related) headlines on social media; and 2) a treatment condition, where they are first asked to indicate how accurate they think a (non-COVID-19 related) news headline is (as part of a pre-test) before making sharing judgments (as in the control). In both studies that will be replicated here, participants will be presented with 15 false and 15 true news content relating to COVID-19, and will be asked either one of the two following questions: "to the best of your knowledge, is the claim in the above headline accurate?" (yes/no) (in our first study); or "if you were to see the above article on social media, how likely would you be to share it?" (on a 6-point Likert scale ranging from "extremely unlikely" to "extremely likely") (in our second study). The original 30 headlines were obtained through a partnership with the Harvard Global Health Institute. False headlines were deemed to be false by authoritative sources such as Snopes and factcheck.org, and other credible sources. Since some of the headlines used in the original study are no longer applicable or valid today, we will replace outdated headlines with relevant ones. A new set of updated headlines will be

provided by the original authors close to the study launch date. A subset of 15 real and 15 fake headlines (out of the 20 real and 20 fake headlines provided by the original authors) will be selected close to the study launch date to ensure that the headlines are as up to date as possible. 15 real and 15 fake headlines will be selected because the original study also used 15 real and 15 fake headlines. We will also explore whether the predicted treatment effect 110arie sas a function of performance on the Cognitive Reflection Test, scientific knowledge, medical maximising-minimising, and political ideology (Democrat versus Republican, continuous scale). Furthermore, participants will be asked two questions specific to the COVID-19 pandemic: "How concerned are you about COVID-19 (the new coronavirus)?" (0 being "not concerned at all" and 100 being "extremely concerned"), and "how often do you proactively check the news regarding COVID-19 (the new coronavirus)?" (1 being "never" and 5 being "very often"). We will also measure the distance from the nearest COVID-19 epicenter (defined as a county with at least 10 confirmed coronavirus cases when the study will be run). In addition to the original study, we also include a rational versus intuitive style decision-making questionnaire, and numeracy measures (using the 3-item Schwartz test and the 4-item Berlin test, as a summed score). These measures will be added to the end of both surveys (as they were not present in the original paper), so that the original design remains intact. Finally, for the treatment condition in our second study, we will record the time elapsed between being exposed to the accuracy nudge and seeing each individual headline. This will take place automatically, without the participants noticing.

- Pennycook_covid_E32j – Direct Replication – Freeman – 528 – Preregistration Draft.docx

Randomization

In our first study, we will only have an accuracy condition, which does not require randomisation. However, the order of presentation as well as the yes/no options will be counterbalanced across participants. In our second study, participants will be randomly assigned to either the treatment or the control condition via the 'randomisation' function in Qualtrics. The 15 true and 15 false headlines will be presented in a random order.

Sampling Plan

Existing Data

Registration prior to creation of data
Explanation of existing data

No response

Data collection procedures

Participants will be recruited through the recruitment company Respondi (for information on the composition of Respondi's US respondent pool see: https://www.respondi.com/EN/access-panel). This is a deviation from the original study, in which participants were recruited through Lucid. We checked that use of an alternative participant source was acceptable with SCORE Co-ordinators. We will use sampling quota to provide a sample matched to the US population on age, gender, ethnicity and geographical location. Participants in the original studies will be excluded as much as possible through asking them whether they have participated in them or not. Participants who indicate not using Facebook or Twitter will also be excluded, as will participants who do not complete the survey. In addition, as an extra attention check, participants will be asked if they responded randomly at any

point during the study or searched for any of the headlines online (e.g. through Google). In line with the original study, we do not intend on excluding individuals who fail such an attention check, but we will conduct exploratory analyses to see if the observed effects remain robust. For details on the study design, we refer to question 8. Participants will be paid approximately £1 for completion of the survey, in line with Respondi's normal payment rate for the length of survey.

No files selected

Sample size

The initial target analytic sample size is 21,013 ratings . If a statistically significant effect is not observed after the first round of data collection, a second round will begin. The second round of data collection will sample an additional 26,443 ratings for a target pooled analytic sample of 47,446 ratings. Each participant rates 30 headlines, so in terms of the number of participants, stage 1 would need 701 , for stage 2 and additional 882 would be needed, for a pooled analytic sample of 1582 (the number is off by 1 due to rounding up to whole participants) . To achieve the target analytic sample, based on the original study, we anticipate the need to collect a larger target sample size to account for failure on the 3 attention checks. The target sample size for recruitment is 770 participants. If necessary, the second round of data collection will sample an additional 960 for a pooled recruited sample size of 1730 participants.

Sample size rationale

Power calculations were done in accordance with the guidelines of the Social Sciences Replication Project (SSRP). The first round of data collection achieves 90% power to detect 75% of the original effect size. The pooled sample, if necessary after testing the effect on the first round of data collection, achieves 90% power to detect 50% of the original effect size. For this replication, the power analysis can be found here:
https://osf.io/b2tzh/?view_only=8°8366d5cdcd409bb8f37ba1067b4db5

Stopping rule

The planned sample size is 21,013 ratings . After achieving that sample, planned analyses will be run. If a significant effect in the hypothesized direction is found, sampling stops. If that significant effect is not found, a second round of data collection will collect data from 26,443 ratings additional observations, for a pooled sample of 47,446 ratings . Each participant rates 30 headlines, so in terms of the number of participants, stage 1 would need 701 , for stage 2 and additional 882 would be needed, for a pooled analytic sample of 1582 (the number is off by 1 due to rounding up to whole participants). Sampling will stop after the second round of data collection regardless of a significant effect. In order to achieve the necessary sample size, we will take into account the attrition rate reported in the original study. For example, in order to achieve a final sample size of N = 856 valid entries for their Study 2, the original authors recruited 1,145 participants who began the study. Out of these, 177 participants indicated not using Facebook or Twitter, and a further 112 did not complete the full study. Similarly, for their Study 1, the original authors recruited 1,143 participants, 192 of which did not indicate using Facebook or Twitter and a further 98 participants did not complete the study, making the final sample size N = 853. We will make use of Twitter or Facebook a criterion for participation in the study in order to prevent the unnecessary attrition of participants on this characteristic. Therefore, in order to achieve our analytical sample targets we will aim to recruit 770 participants for our first study (in order to match the original in variance as closely as possible), and 950 for our second study. Our quota-based sampling approach will allow real-time identification of participants who meet screening and exclusion criteria (via the Qualtrics survey

platform), but not those who fail attention checks. We will continue to recruit further if initial analysis reveals that we have not reached the intended analytic sample size with our target recruitment sample size.

Variables

Manipulated variables

In our study 2, the treatment condition will be shown the same false and true headlines about COVID-19 as the control condition and the accuracy condition, the only difference being that the treatment condition is first asked to rate the accuracy of a non-COVID-related headline ("To the best of your knowledge, is the claim in the above headline accurate?" (yes/no)) prior to being asked about sharing intent, which is framed as being for a pretest (in line with Pennycook, Epstein et al. (2019)). For further details, we refer to question 8.

No files selected

Measured variables

Participants will be presented with 15 false and 15 true news content relating to COVID-19, and will be asked either one of the two following questions: (in the treatment and control conditions in our study 2) "if you were to see the above on social media, how likely would you be to share it?" (on a 6-point Likert scale ranging from "extremely unlikely" to "extremely likely"); or (in our study 1) "to the best of your knowledge, is the claim in the above headline accurate?" (yes/no). In terms of covariates, we will explore whether the predicted treatment effect 112arie sas a function of performance on the Cognitive Reflection Test (Frederick, 2005; Toplak et al., 2011), consisting of a reworded version of the original 3-item test and 3 items from a non-numeric version, excluding the "hole" item (Thomson & Oppenheimer, 2016) ; scientific knowledge, a measure of general background knowledge for scientific issues – consisting of 17 questions about basic science facts (e.g., "Antibiotics kill viruses as well as bacteria", "Lasers work by focusing sound waves") (McPhetres & Pennycook, 2020) ; the medical maximising-minimising scale (Scherer et al., 2016) , which measures the extent to which people are either "medical maximisers" who tend to seek health care even if for minor issues or, rather, "medical minimisers" who tend to avoid health care unless absolutely necessary; political ideology (Democrat versus Republican, continuous scale); and 2 measures that were not part of the original study: rational versus intuitive style decision-making (Hamilton et al., 2016) ; and numeracy (using the 3-item Schwartz test and the 4-item Berlin test as a summed score (Cokely et al., 2012; Schwartz et al., 1997)).

No files selected

Indices

Indices are described in detail in the "analysis plan" section.

No files selected

Analysis Plan

Statistical models

For the purposes of SCORE, to test H*, H1 and H2, in line with the original paper, we will test whether the treatment condition differs from the control condition by testing for an interaction

effect between condition (0=control, 1=treatment) and news type (0=false, 1=true). We will then test for a simple effect of both news type and condition for each of the two types of news (the treatment effect is predicted to be larger for fake news). In line with the original paper, analyses will first be conducted at the level of the rating, using linear regression with robust standard errors clustered on participants and headline. In addition, we will also conduct independent samples t-tests and ANOVAs to check for between-group differences in sharing intent between the treatment and control groups. We will also perform an item-level analysis to look at whether increasing individuals' attention to accuracy yields larger changes in sharing intentions for different headline types. For each headline, we examine how the effect of the treatment on sharing (i.e., average sharing intention in treatment minus average sharing intention in the control) varies based on the average accuracy rating given to that headline by participants in the "accuracy" condition. All analyses will be conducted in R. Please see attached code (study 1 and study 2) and test data (for study 1 and study 2).

- Pennycook et al (2020) – study 1 code.do
- Pennycook et al. (2020) – study 2 code.do
- Pennycook et al. (2020) – test data (study 1).csv
- Pennycook et al. (2020) – test data (study 2).csv

Transformations

In line with the original paper, analyses will be conducted at the level of the rating, using linear regression with robust standard errors clustered on participants and headline. Sharing intentions will thus be rescaled so that 1 on the 6-point Likert scale is 0 and 6 on the 6-point Likert scale is 1.

Inference criteria

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal hypothesis test (H*). For this study a significant positive interaction between treatment and headline veracity predicting likelihood to share will be considered a successful replication (outcome of the F test). The pattern of results is that, in the control condition, there was not a significant difference in likelihood to share between true and false headlines, but in the treatment condition, participants were significantly more likely to share a true headline compared to a false headline.

Data exclusion

Participants who indicate not using Facebook or Twitter will also be excluded, as will participants who do not complete the survey. Due to quota-based sampling participants who select 'other' or 'prefer not to say' as gender will not be included in the final sample for analysis. Including a non-binary catergory in quotas for gender is impractial given the sample size (the estimated proportion of the US population is too small). Following from the original study, we will also treat gender as a binary covariate. In addition, as an extra attention check, participants will be asked if they responded randomly at one point during the study or searched for any of the headlines online (e.g. through Google). In line with the original study, we do not intend on excluding individuals who fail such an attention check, but we will conduct exploratory analyses to see if the observed effects remain robust.

Missing data

Missing data will be removed through casewise removal. All incomplete responses will be excluded

from the analysis, meaning that if one response is missing, they will be removed from the data set.

Exploratory analysis

We are conducting 2 studies, in line with Pennycook, McPhetres et al. (2020) (the original study), which is essential for replicating the full study. The first study, which will include only 1 condition (the accuracy condition) will be used to calculate the average reported accuracy of all 30 headlines related to COVID-19 that are shown to participants in the study. If there is a significant effect for the treatment condition in the second study, we will explore the treatment effect by level of the covariates (mentioned above). We will also conduct additional t-tests/ANOVAs in addition to the analyses performed in the original paper, as an extra robustness check.

Other

Other

All data and analysis scripts will be made available on the OSF. This study deviates from the original study in a number of ways: first, the 30 news headlines about COVID-19 used in the original study were obtained through a partnership with the Harvard Global Health Institute. False headlines were deemed to be false by authoritative sources such as Snopes and factcheck.org, and other credible sources. Since some of the headlines used in the original study are no longer applicable or valid today, we will replace outdated headlines with relevant ones. A new set of updated headlines will be provided by the original authors close to the study launch date. A subset of 15 real and 15 fake headlines (out of the 20 real and 20 fake headlines provided by the original authors) will be selected close to the study launch date to ensure that the headlines are as up to date as possible. 15 real and 15 fake headlines will be selected because the original study also used 15 real and 15 fake headlines. Second, our first study is a minor deviation from the original study (Study 1 in the original paper), as it includes only 1 condition (the accuracy condition). The results from this study (specifically: average accuracy scores for true and false headlines about COVID-19) were used for further analyses in the original paper's Study 2 (which we are replicating). Third, in addition to the original study's covariates, we also include a rational versus intuitive style decision-making questionnaire, and a numeracy measure (using the 3-item Schwartz test and the 4-item Berlin test, as a summed score). These two measures will be added to the end of the survey (as they were not present in the original paper), so that the original study design remains intact.

Preregistrazione di Meier, M. E. (2019). Is there a positive association between working memory capacity and mind wandering in a low-demanding breathing task? A preregistered replication of a study by Levinson, Smallwood, and Davidson (2012). *Psychological Science*, *30*(5), 789-797.

Persistence of Thought Replication (#2803)

Author(s)
Matt Meier (Western Carolina University) - mmeier@wcu.edu
Adam Lyons (Western Carolina University) - ajlyons1@catamount.wcu.edu

Created: 02/02/2017 07:25 AM (PT)
Public: 01/19/2018 11:37 AM (PT)

1) Have any data been collected for this study already?

No, no data have been collected for this study yet

2) What's the main question being asked or hypothesis being tested in this study?

Can we replicate the Levinson, Smallwood, and Davidson's (2012) Experiment 2 finding that higher-WMC subjects mind wandered (as measured by experimenter-controlled probes) more than lower-WMC subjects while completing a breath awareness task. Furthermore, we want to estimate the size of this effect with more precision than the original accomplished.

3) Describe the key dependent variable(s) specifying how they will be measured.

The key dependent variable in this study is the correlation between performance on working memory capacity (WMC) measurement tasks (i.e., complex span tasks) and the percentage of probes that subjects rate as 4 or higher (on a 6-point Likert scale). Scores of 4 or higher are considered task unrelated thoughts (TUTs). As in Levinson et al., we will measure WMC with the operation span task. In addition, we will measure WMC with a symmetry span task. The dependent variable in the complex span tasks will the number of memoranda recalled in the correct serial order. We will also form a WMC composite by converting these raw scores (one from each complex span task) into Z scores (means and standard deviations from the same sample) and averaging them.

4) How many and which conditions will participants be assigned to?

In this quasi-experimental study there is only one condition to which all subjects will be assigned.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will conduct correlation analyses between the individual complex span tasks and the percentage of TUTs (i.e., mind wandering) and a correlation between the WMC composite and percentage of TUTs. In the original work, Operation span task scores were square root transformed to adjust for skewness. We will conduct all analyses with span scores transformed and without transformation.

6) Any secondary analyses?

We will also analyze relations between complex span tasks (and the WMC composite) and percentage of TUTs in a 6 minute baseline task, a 20 minute breath counting task, and overall percentage of task unrelated thoughts across the three breathing tasks.

In the awareness portion of the task we will examine the number of self-caught mind wandering reports. We will test if these relate to our complex span measures. We will look at self-caught mind wandering as raw counts and as a ratio with probe-caught mind wandering within the awareness task and the overall rate of mind wandering across the three breathing tasks.

7) **How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how thenumber will be determined.**

Because our goal with this study is a precise estimate, we will stop data collection at the end of a semester in which we have at least 220 subjects with 2 complex span tasks that meet the 85% processing criterion.

8) **Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)**

We will follow Levinson et al. and convention by excluding complex span task scores where subjects do not make an 85% correct criterion in the processing portion of the task. Without looking at their data, we will also exclude subjects from analysis who experimenters judge to be noncompliant with task instructions (e.g., sleeping, obviously not following task instructions).

Levinson et al. used a community sample for this study who received payment for participation. In this replication, we will use university students who receive credit towards a course requirement.

Preregistrazione di Tybur, J. M., Jones, B. C., DeBruine, L. M., Ackerman, J. M., & Fasolt, V. (2020). Preregistered direct replication of "sick body, vigilant mind: the biological immune system activates the behavioral immune system". *Psychological Science*, *31*(11), 1461–1469.

Study Information

Title

Preregistered replication of “Sick body, vigilant mind: The biological immune system activates the behavioral immune system”

Authors

Publication Information

In order to complete this registration your study must be granted an "in-principle acceptance" from a

journal that offers Registered Reports.

I confirm this study has been granted in-principle acceptance.

Journal title

Psychological Science

Date of in-principle acceptance

2019-03-25

Manuscript

Attach manuscript

- Illness Recency Proposal_IPA.docx

Other

Attach any supporting documents

- SBVM_rep_v4.html
- SBVM_rep_v4.Rmd

Other information

No response

Preregistrazione di Halfmann, E., Bredehöft, J., & Häusser, J. A. (2020). Replicating roaches: A preregistered direct replication of Zajonc, Heingartner, and Herman's (1969) social-facilitation study. *Psychological Science*, *31*(3), 332–337.

Study Information

Title

Cockroaches 2.0 - a replication of Zajonc et al., 1969.

Research Questions

This pre-registered study seeks to replicate Zajonc's famous social facilitation effect (Zajonc, Heingartner, & Herman, 1969; JPSP). Zajonc and colleagues employed two studies in which they observed the behavior of cockroaches under two different types of social treatments, coactions and

audience. They assessed the running times of the cockroaches needed to traverse a runway versus a maze. The results of study 1 showed that the mere presence of audience enhances the performance in an easy task (runway) and impairs it in a difficult one (maze). We want to examine whether this interaction effect replicate. Therefore, we concentrate on the interaction of task difficulty and presence of audience (part of study 1, leaving out the coaction conditions).

Hypotheses

We expect a sig. interaction of task difficulty and presence of audience on running time. More precisely, we expect that cockroaches finish an easy task (runway) faster if an audience is present compared to running without an audience. In contrast, their performance will be impaired in a difficult task (maze) if an audience is present compared to running without audience.

Sampling Plan

Existing Data

Registration prior to creation of data
Explanation of existing data

The pre-registration is uploaded prior to the beginning of data collection.

Data collection procedures

Participants will be selected from one discrete colony of death's head cockroaches (blaberus caniifer). Only female adult individuals are selected by trained zookeeper. For at least one week prior to the testing all test subjects were housed in individual opaque boxes. They were maintained in quarters with 12h of constant daylight and a constant temperature of 23 degrees Celsius. The cockroaches were fed a diet of peeled apples.

No files selected

Sample size

Our target sample is 120 cockroaches that will be randomly assigned to the conditions of a 2 (audience vs. non audience) x 2 (runway vs. maze). Additionally, a group of 40 cockroaches is serving as audience.

Sample size rationale

Unfortunately, the data presented by Zajonc et al. (1969) do not allow for calculating effect sizes (no effect sizes, only means, no SDs or SEs, no F-value for the critical interaction, no post-hoc contrast). Zajonc et al. (1969) used n = 10 cockroaches in each condition, with each cockroach running 10 trials (in order to reduce error variance). We decided to triple the sample size of the original study (N = 120), allowing us to detect medium sized effects (f =.3), with α = .05 and a power of .9.

Stopping rule

Data collection will be terminated when the target sample size of N = 120 is achieved.

Variables

Manipulated variables

All roaches have to complete a task by running from a starting box through a runway (or maze) into a goal box. We will manipulate difficulty of the task as between factor with two levels: difficult and easy. A straight runway served as easy task and a maze as difficult task. The audience was manipulated by having the subjects perform the tasks either in the presence of 40 cockroaches, separated into four boxes, or alone. The boxes were placed inside a plexiglass cube, with sides directly contiguous with the walls of the runways and aligning air holes allowing the transmission of olfactory cues.

No files selected

Measured variables

1. As main dependent variable, we will measure the running time each subject needs to finish the task. Measurement started with the cockroach leaving the starting box and ended when the cockroach entered the goal box completely. 2. We will also measure the starting latency, which is the time the cockroach needs to leave the starting box after the door separating the starting box from the runway/maze is opened.

No files selected

Indices

Since each subject will complete the assigned task ten times, running time and starting latencies are averaged over all trials.

No files selected

Design Plan

Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
Blinding

- No blinding is involved in this study.

Study design

A two-factorial between subject design will be used, with task difficulty (2 levels: runway vs. maze) and presence of audience (2 levels: present vs. running alone) as between factors.

No files selected

Randomization

Participants will be randomly picked by the zookeepers and assigned to one of four conditions.

Analysis Plan

Statistical models

As we are interested in the relationship between task difficulty and presence of audience, we will look into this interaction using a two-way ANOVA and planed contrast to estimate the simple effects.

No files selected

Transformations

N/A

Follow-up analyses

Planed contrasts will be conducted to test simple main effects.

Inference criteria

We will use the standard p < .05 criteria

Data exclusion

If a subject does not leave the starting box within 3 minutes the trial will be aborted and restarted.

Missing data

We will only analyze data from subjects who completed at least eight trials successfully.

Exploratory analysis

No response

Scripts

Upload an analysis script with clear comments

No files selected

Other

Other

A detailed description of the procedure as well as the construction plan of the apparatus can be found in the original study by Zajonc at al. (1969) (Study 1, leaving out the coaction conditions). We replicate the apparatus and the general procedure as close as possible.

Preregistrazione di Walmsley, J., & O'Madagain, C. (2020). The worst-motive fallacy: A negativity bias in motive attribution. *Psychological Science*, *31*(11), 1430–1438.

The Nature of the Effect

Verbal description of the effect I am trying to replicate

In Walmsley and O'Madagain "The Worst Motive Fallacy" we found that participants were

significantly inclined to expect an agent in a story to act on the worst of two motives we described the agent as having. We want to replicate this in the following study.

It is important to replicate this effect because

In our replication, we will use 12 vignettes instead of just the 4 from our original study (we will replicate the effect of the 4, and add 8 more). This will increase the diversity of the vignettes, and also allow us to include vignette as a random effect in the model for the study (a random effect with four levels is not optimal). Also, the effect we found before was significant but not highly. If the effect is real, it should replicate.

The effect size of the effect I am trying to replicate is

n/a. effects reported through model comparison

The confidence interval of the original effect is

n/a. effects reported through model comparison

The sample size of the original effect is

323

Where was the original study conducted? (e.g., lab, in the field, online)

online - qualtrics

What country/region was the original study conducted in?

USA

What kind of sample did the original study use? (e.g., student, Mturk, representative)

Mturk

Was the original study conducted with paper-and-pencil surveys, on a computer, or something else?

Computer

Designing the Replication Study

Are the original materials for the study available from the author?

yes

I know that assumptions (e.g., about the meaning of the stimuli) in the original study will also hold in my replication because

They are identical

Location of the experimenter during data collection

n/a

Experimenter knowledge of participant experimental condition

n/a

Experimenter knowledge of overall hypotheses

n/a

My target sample size is

320*3=980

The rationale for my sample size is

We are using three times as many stimuli as in the original, so we want to increase the sample size sufficiently to include vignette as a random effect, hence multiplying the sample size by 3.

Documenting Differences between the Original and Replication Study

The similarities/differences in the instructions are

Exact

The similarities/differences in the measures are

Close

The similarities/differences in the stimuli are

Exact

The similarities/differences in the procedure are

Exact

The similarities/differences in the location (e.g., lab vs. online; alone vs. in groups) are

Exact

The similarities/difference in remuneration are

Exact

The similarities/differences between participant populations are

Close

What differences between the original study and your study might be expected to influence the size and/or direction of the effect?

We are replicating the original four vignettes but also eight more, to increase the plausibility of our conclusion (that the effect is independent of the context). The additional vignettes may affect the outcome. We are also including vignette as a random effect in the model, which is a sound approach now that we have twelve vignettes - this may also affect the outcome.

I have taken the following steps to test whether the differences listed in the previous question will influence the outcome of my replication attempt

None. We have made the prediction that our effect is independent of context - therefore we should be able to find the effect across new vignettes with the same basic structure. If we pilot these vignettes to "see if they work", this would bias our selection of vignettes unreasonably in our favor.

Analysis and Replication Evaluation

My exclusion criteria are (e.g., handling outliers, removing participants from analysis)

Participants are given three very easy 'attention questions' at the outset of the study, and those who do not answer all three correctly are excluded on the assumption that they are not paying attention. Participants who do not complete the study are excluded.

My analysis plan is (justify differences from the original)

Include random effect of vignette, whereas vignette was included as a fixed effect in our original study. The inclusion of vignette as a random effect is to take a more conservative approach to the claim that the effect is not dependent on the vignette chosen.

A successful replication is defined as

The study will replicate successfully if, for the original vignettes as well as new vignettes with the same basic structure, participants significantly expect the agent in the vignette to act on the worst of two competing motives she has.

Preregistrazione di De Wilde, M., Casini, A., Bernard, P., Wollast, R., Klein, O., & Demoulin, S. (2020). Two preregistered direct replications of "objects don't object: evidence that self-objectification disrupts women's social activism". *Psychological Science*, *31*(2), 214–223.

Studio-replica 1:

A. Hypotheses - Essential elements

Description of essential elements

Describe the (numbered) hypotheses in terms of directional relationships between your (manipulated or measured) variables.

Since the current project consist in a replication attempt, our hypotheses are exactly the same as in the paper of Rachel Calogero “Objects Don't Object: Evidence That Self-Objectification Disrupts Women's Social Activism” (2013). H1: self-objectification is negatively related to social activism H2: self-objectification is positively related to gender-specific system justification H3: gender-specific system justification is negatively related to social activism H4: gender-specific system justification mediated the negative relationship between trait self-objectification and gender-based social activism

For interaction effects, describe the expected shape of the interactions.

N/A

If you are manipulating a variable, make predictions for successful check variables or explain why no manipulation check is included.

N/A

Recommended elements

Recommended elements

A figure or table may be helpful to describe complex interactions; this facilitates correct specification of the ordering of all group means.

- Papp & Erchull, 2016.pdf
- Calogero, Tylka, Donnelly, McGetrick, & Leger, 2017.pdf
- Calogero, 2013.pdf

For original research, add rationales or theoretical frameworks for why a certain hypothesis is tested.

N/A (all the rationales or theoretical frameworks are presented in Calogero's works). The reason why we attempt a preregistered direct replication is because we already lead 3 indirect close replication of the interest with high power (.99) and larger samples (n1 = 194; n2 = 233 & n3 = 172) than the original study (n=50).

If multiple predictions can be made for the same IV-DV combination, describe what outcome would be predicted by which theory.

N/A

B. Methods - Essential elements

Description of essential elements

Design

List, based on your hypotheses from section A: Independent variables with all their levels a. whether they are within- or between-participant b. the relationship between them (e.g., orthogonal, nested).

N/A

List dependent variables, or variables in a correlational design

Conceptual IV: Sexual self-objectification Conceptual Mediator: Gender-specific system justification Conceptual DV: Social activism

Third variables acting as covariates or moderators.

Ethnicity

Planned Sample

If applicable, describe pre-selection rules.

The sample will be as similar as possible to the one of Calogero, Study 1 (2013): Calogero' sample: Mage=18.65; SD = 2.11 years Our sample pre-selection rules: Age: between 18 and 25 years Nationality: US Sex: Women First language: English

Indicate where, from whom and how the data will be collected.

data will be collected on Prolific by Matthias De Wilde via an online survey

Justify planned sample size

Expected n = 90 for a statistical power of .99 (calculate based on the effect size of Calogero, 2013, Study1)

If applicable, you can upload a file related to your power analysis here (e.g., a protocol of power analyses from G*Power, a script, a screenshot, etc.).

- Power analysis.docx

Describe data collection termination rule.

When we obtain a n = 90, the study will be interrupted

Exclusion Criteria

Describe anticipated specific data exclusion criteria. For example: a) missing, erroneous, or overly consistent responses; b) failing check-tests or suspicion probes; c) demographic exclusions; d) data-based outlier criteria; e) method-based outlier criteria (e.g. too short or long response times).

a. Failing attention check questions (e.g., "if you read this item, answer "3 = neither agree, nor disagree") b. Data-based outlier criteria (specifically: being a statistical outlier on at least two variable) c. Failing to understand the instructions of the SOQ = if the sum of the non-observable items score OR the sum of the observable items score is higher than 25

Procedure

Describe all manipulations, measures, materials and procedures including the order of presentation and the method of randomization and blinding (e.g., single or double blind), as in a published Methods section.

The preview of the study is available following this link: https://uclpsychology.co1.qualtrics.com/jfe/preview/SV_4MmwfTIo5Vi7eLj?Q_SurveyVersionID=current&Q_CHL=preview Our procedure is exactly similar to the one of Calogero. Indeed, she fairly agree to provide her consent to check if our design was a proper replication of her study (cfr. Mail) The three following measure will be presented in a counterbalanced order: a. The Self-Objectification Questionnaire (SOQ; Noll & Fredrickson, 1998). Participants will be instructed to rank the 10 attributes in order of impact on their physical self-concept, from 1 (greatest impact on my physical self-concept) to 10 (least impact on my physical self-concept). The same rank could not to be assigned to more than 1 attribute. We change the anchor as compare to Calogero who coded 0 (least impact on my physical self-concept) to 9 (greatest impact on my physical self-concept) to be consistent with the traditional use of the scale. Moreover Calogero reported to found a correlation of r = -.97 between the sum of the "non-observable aspect" and the sum of the "observable aspect" of the SOQ. Such a correlation is impossible since participant should not have assigned the same rank to different attribute. This correlation should systematically be of r = -1 if participants correctly

answer the questionnaire. b. Gender-specific system justification will be measured with eight items of Jost & Kay (2005) reflecting the extent to which participants endorsed the current state of gender relations ((a) “In general, relations between men and women are fair,” (b) “The division of labor in families generally operates as it should,” (c) “Gender roles need to be radically restructured,” (d) “For women, the United States is the best country in the world to live in,” (e) “Most policies relating to gender and the sexual division of labor serve the greater good,” (f) “Everyone (male or female) has a fair shot at wealth and happiness,” (g) “Sexism in society is getting worse every year,” and (h) “Society is set up so that men and women usually get what they deserve.” Participants were asked to indicate the strength of agreement or disagreement with each of these items on a 9-point scale. Responses were coded in such a way that agreement with Items a, b, d, e, f, and h and disagreement with Items c and g resulted in higher scores on gender-specific system justification. The overall index was α= .65 in Jost & Kay’ study (2005) and α= .85 in Calogero’s study. Responses were made using scales from 1 (strongly disagree) to 9 (strongly agree). Items were averaged to create system-justification scores (α = .85). Scores ranged from 1 to 9, with higher scores indicating greater justification of the gender status quo. c. A set of feminist-activism items derived from prior research will be used to measure gender-based social activism (Stake, Roades, Rose, Ellis, & West, 1994). Participants rated eight items assessing the extent to which they had participated in various types of social activism in the area of gender equality during the previous 6 months; responses were made on scales from 1 (never) to 7 (all the time). The eight items covered the following types of activism: discussing issues related to gender equality with friends or colleagues (in person or online—e.g., through e-mail, Facebook, Twitter, or MySpace); attending meetings, conferences, or workshops on gender-equality issues; signing a petition (in person or online) in support of women’s rights and gender equality; circulating a petition (in person or online) related to women’s rights or gender equality; handing out fliers related to women’s-rights issues or gender equality; attending demonstrations, protests, or rallies related to women’s rights or gender equality; working for women’s rights campaigns (e.g., fund-raising); and acting as a spokesperson for a particular gender-equality issue. Ratings for all items were averaged to create social-activism scores (α = .92). Scores ranged from 1 to 7, with higher scores indicating more participation in gender-based social activism.

Recommended elements

Recommended elements

Procedure

Set fail-safe levels of exclusion at which the whole study needs to be stopped, altered, and restarted. You may pre-determine what proportion of excluded participants will cause the study to be stopped and restarted.

N/A

If applicable, you can upload any files related to your methods and procedure here (e.g., a paper describing a scale you are using, experimenter instructions, etc.)

No files selected

C. Analysis plan - Essential elements

Confirmatory Analyses

Describe the analyses that will test the first main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

We will test a simple mediation model using PROCESS (Model 4; Hayes, 2013) to examine the direct and indirect effect of self-objectification on social activism through system justification. Self-objectification will be entered as the predictor (X), social activism will be entered as the criterion (Y), and system justification (M) will be entered as the mediating variable. Significance of indirect paths was assessed using 95% bias-corrected and accelerated confidence intervals with 10000 bootstrap resamples.

the statistical technique;

idem

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

IV = Self-objectification Mediator = Gender-specific system justification DV = Social activism

rationale for each covariate used, if any;

N/A

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

if we found an effect we will use Bayesian statistics

Second Prediction

Describe the analyses that will test the second main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Third Prediction

Describe the analyses that will test the third main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Fourth Prediction

Describe the analyses that will test the fourth main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Further Predictions

Describe the analyses that will test any further (main) predictions from the

hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Recommended elements

Recommended Elements

Specify contingencies and assumptions, such as:

Method of correction for multiple tests.

No response

The method of missing data handling (e.g., pairwise or listwise deletion, imputation, interpolation).

No response

Reliability criteria for item inclusion in scale.

No response

Anticipated data transformations.

No response

Assumptions of analyses, and plans for alternative/corrected analyses if each assumption is violated.

No response

Optionally, upload any files here that are related to your analyses (e.g., syntaxes, scripts, etc.).

No files selected

Final questions

Has data collection begun for this project?

No, data collection has not begun

If data collection has begun, have you looked at the data?

No

The (estimated) start and end dates for this project are

december 2018 - december 2019

Any additional comments before I pre-register this project

No response

Studio-replica 2:

A. Hypotheses - Essential elements

Description of essential elements

Describe the (numbered) hypotheses in terms of directional relationships between your (manipulated or measured) variables.

Based on 4 close replications of Rachel Calogero's paper "Objects Don't Object: Evidence That Self-Objectification Disrupts Women's Social Activism" (2013), we led a Preregistered Direct Replication (PDR1) of this work. Results of PDR1 confirms that the effect observed by Calogero has been overestimated. We now have explicit a priori evidence that the effect she observed was, to the least, overestimated (4 studies and PDR1). However, we conducted the PDR1 by calculating the power needed based on the (suspected to be overestimated) effect size she reported in her paper. Thus, we conducted a second PDR (PDR2) with 2.5 times the sample size of Calogero following the recommendation of D. Stephen Lindsay (Editor, Psychological Science) Authors: Matthias De Wilde1, Annalisa Casini1, Philippe Bernard2, Robin Wollast2, Olivier Klein2, and Stéphanie Demoulin1 Affiliation : 1Université catholique de Louvain (UCL), Belgium 2Université libre de Bruxelles (ULB), Belgium H1: self-objectification is negatively related to collective action intentions H2: self-objectification is positively related to gender-specific system justification H3: gender-specific system justification is negatively related to collective action intentions H4: gender-specific system justification mediated the negative relationship between trait self-objectification and gender-based collective action intentions More specifically, we will try to replicate this model described in Calogero (2013)

For interaction effects, describe the expected shape of the interactions.

N/A

If you are manipulating a variable, make predictions for successful check variables or explain why no manipulation check is included.

N/A

Recommended elements

Recommended elements

A figure or table may be helpful to describe complex interactions; this facilitates correct specification of the ordering of all group means.

- OSF Pre-Registration Calogero_PDR2.pdf
- model Calogero (2013).png

For original research, add rationales or theoretical frameworks for why a certain hypothesis is tested.

N/A (all the rationales or theoretical frameworks are presented in Calogero (2013). We conducted this preregistered direct replication based on four unpublished replication attempts and one preregistered direct replication following the recommendations of D. Lindsay, editor in Psychological Science)

If multiple predictions can be made for the same IV-DV combination, describe what outcome would be predicted by which theory.

N/A

B. Methods - Essential elements

Description of essential elements

Design

List, based on your hypotheses from section A: Independent variables with all their levels a. whether they are within- or between-participant b. the relationship between them (e.g., orthogonal, nested).

Here is the link leading to the preview of the Survey: https://uclpsychology.co1.qualtrics.com/jfe/preview/SV_4MmwfTIo5Vi7eLj?Q_SurveyVersionID=current&Q_CHL=preview The scales used in the present study to measure the three variables of the original model will be identical to the ones used in Calogero's study except for minor improvements described in the "method" and implemented with her approval (available on OSF). Two attentional check questions were included in the survey to make sure that participants were carefully completing the survey (e.g., "If you are reading this question, answer "sometimes" so that we can check if you have read the questions"). Calogero confirmed our material was similar and the improvements were relevant before collecting the data of the present study. All the participants will answer to three questionnaires (plus the demographic variables): Independent variables: Self-Objectification: Self-Objectification Questionnaire (Noll & Fredrickson, 1998) was used to measure self-objectification. Participants had to rank from 1 to 10 (From 0 to 9 in the original study) the importance of 10 physical attributes for their self-concept. Of these physical attributes, five were observable (e.g., weight) and five were not (e.g., health). Scores were calculated by subtracting the sum of the ranks of the unobservable attributes from the sum of the ranks of the appearance-based attributes. Score ranged from -25 to 25 and higher scores indicated greater self-objectification. A correlation of -1 between the mean of the ranks of the observable attributes was found in PDR1

(confirming that participant understood the instructions correctly).

List dependent variables, or variables in a correlational design

Gender-Specific System Justification: Gender-Specific System Justification Scale (Jost & Kay, 2005) was used to measure GSSJ. Participants were asked to indicate their level of agreement using a 9-point Likert scale ranging from 1 (“Strongly disagree”) to 9 (“Strongly agree”) on 9 items. Responses were coded in such a way that agreement with items a, b, d, e, f, and h and disagreement with items c and g resulted in higher scores on GSSJ. A sample item is “In general, relations between men and women are fair”. In PDR1, this scale presented a good reliability (α = .85). Collective Action Intentions: The scale used to assess the intention to engage in collective actions, adapted by Calogero from Stake, Roades, Rose, Ellis, and West (1994) in the original study, was used to measure collective action intentions. Participants were asked to indicate the frequency with which they had participated in 8 different types of actions considered as social activism in the area of gender equality during the 6 months preceding their participation to the survey using a 7-point Likert scale ranging from 1 (“Never”) to 7 (“Always”). We duplicated two items of the original scale in two alternatives (“Sign an online petition for women's rights and gender equality” vs. “Sign a petition in the street for women's rights and gender equality”). In PDR1, this scale presented a good reliability (α = .90).

Third variables acting as covariates or moderators.

Calogero did not mention any covariate in the method section of her paper (2013). We intent to provide the full data on OSF for people interested in conducting the analyses with the possible covariates (e.g., age, ethnicity).

Planned Sample

If applicable, describe pre-selection rules.

The sample will be as similar as possible to the one of Calogero, Study 1 (2013): Calogero’s sample (Mage= 18.65; SD = 2.11 years). Our sample pre-selection rules: Age: between 18 and 25 years Nationality: US and British (sample of PDR1 was composed of women of US nationality) Sex: Women First language: English Not having participated to PDR1

Indicate where, from whom and how the data will be collected.

Data will be collected on Prolific Academic by Matthias De Wilde using an online survey crafted with Qualtrics. Prolific Academic (i.e., a crowdsourcing platform dedicated to academic purposes) and that has been demonstrated to be a reliable and cost-effective source of high-quality and representative data, for multiple research purposes, in and outside the behavioral sciences (Peer, Brandimarte, Samat, & Acquisti, 2017). We still anticipate a usual loss of data of average 10% for uncompleted answers (Peer, Brandimarte, Samat, & Acquisti, 2017, p. 155). Participants will be compensated £0.5 for completing the study, which lasted on average 5.22 minutes in PDR1.

Justify planned sample size

In the file "Power Analyses PDR1" are repported the power analyses we based on for PDR1. The sample size of the original study was 50. We targeted a minimum sample size of 80 based on a priori power analysis (power of .98 with a 95% confidence level) relying on the effects size repported by Calofgero (2013). However, we had good evidence that the effect reported in Calogero (2013) was overestimated. We thus decided to increase the sample for PDR2. We followed the recommendation formulated by Schönbrodt & Perugini (2013) to maintain stability in correlational studies and to

avoid noisiness in measures (e.g., https://www.nicebread.de/at-what-sample-size-do-correlations-stabilize/), and using the perspective of the small telescope argument of Simonsohn (2015), we estimated a sample size of at least 2.5 times the one of the original study (i.e., 50 women). That is a sample size of 160 women (considering the possible loss of data from Prolific). Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? Journal of Research in Personality, 47, 609-612. doi:10.1016/j.jrp.2013.05.009 Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). Downloaded from http://science.sciencemag.org/ on May 9, 2019.

If applicable, you can upload a file related to your power analysis here (e.g., a protocol of power analyses from G*Power, a script, a screenshot, etc.).

- Power Analysis PDR1.pdf

Describe data collection termination rule.

The data collection will be ended when we achieve 160 complete responses on Prolific Academic.

Exclusion Criteria

Describe anticipated specific data exclusion criteria. For example: a) missing, erroneous, or overly consistent responses; b) failing check-tests or suspicion probes; c) demographic exclusions; d) data-based outlier criteria; e) method-based outlier criteria (e.g. too short or long response times).

a. Failing attention check questions (e.g., "if you read this item, answer "3 = neither agree, nor disagree") b. Data-based outlier criteria Detection: - Univariate outliers: we plan to identify as outliers, data that are outside of the following range: median +/-3MAD as suggested by Leys et al. (2013). - Multivariate outliers: we shall examine bivariate relations between Self-Objectification on the one hand and System-Justification and Intentions to engage in collective action on the other hand. We shall use the minimum covariance approach with a breakdown point of .25 as recommended by Leys et al. (2013, 2019). Handling: - We shall report the correlations between our variables with all data included and whether outlier exclusion changes the conclusions of the analysis. Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764–766. DOI: https://doi.org/10.1016/j.jesp.2013.03.013 Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. Journal of Experimental Social Psychology, 74, 150–156. DOI: https://doi.org/10.1016/j.jesp.2017.09.011 Leys, C., Delacre, M., Mora, Y., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. Revue internationale de psychologie sociale, 32, 1-10. issn:0992-986X Failing to understand the instructions of the SOQ = if the sum of the non-observable items score OR the sum of the observable items score is higher than 25.

Procedure

Describe all manipulations, measures, materials and procedures including the order of presentation and the method of randomization and blinding (e.g., single or double blind), as in a published Methods section.

Each of the scales was presented to participants in a randomized order to reduce common method variance problems (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Here is the link leading to the preview of the Survey:

https://uclpsychology.co1.qualtrics.com/jfe/preview/SV_4MmwfTIo5Vi7eLj?Q_SurveyVersionID=current&Q_CHL=preview

Recommended elements

Recommended elements

Procedure

Set fail-safe levels of exclusion at which the whole study needs to be stopped, altered, and restarted. You may pre-determine what proportion of excluded participants will cause the study to be stopped and restarted.

N/A

If applicable, you can upload any files related to your methods and procedure here (e.g., a paper describing a scale you are using, experimenter instructions, etc.)

- OSF Pre-Registration Calogero_PDR2.pdf

C. Analysis plan - Essential elements

Confirmatory Analyses

Describe the analyses that will test the first main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

Our procedure is exactly similar to the one of Calogero except for minor improvement fully reported in the manuscript. Indeed, Calogero agreed to provide her consent to check if our survey was an exact replication of her own study (follow-up of mail available on request). The three following measures will be presented in a counterbalanced order: a. The Self-Objectification Questionnaire (SOQ; Noll & Fredrickson, 1998). Participants will be instructed to rank the 10 attributes in order of impact on their physical self-concept, from 1 (greatest impact on my physical self-concept) to 10 (least impact on my physical self-concept). The same rank can not be assigned to more than 1. We change the anchor as compared to Calogero who coded 0 (least impact on my physical self-concept) to 9 (greatest impact on my physical self-concept) to be consistent with the traditional use of the scale. Moreover Calogero reported to find a correlation of r = -.97 between the sum of the "non-observable aspect" and the sum of the "observable aspect" of the SOQ. Such a correlation is impossible since participants should not have assigned the same rank to different attributes. This correlation should systematically be of r = -1 if participants correctly answer the questionnaire. Furthermore, the scores should range from −25 to +25. b. Gender-specific system justification will be measured with eight items of Jost & Kay (2005) reflecting the extent to which participants endorsed the current state of gender relations ((a) "In general, relations between men and women are fair," (b) "The division of labor in families generally operates as it should," (c) "Gender roles need to be radically restructured," (d) "For women, the United States is the best country in the world to live in," (e) "Most policies relating to gender and the sexual division of labor serve the greater good," (f) "Everyone (male or female) has a fair shot at wealth and happiness," (g) "Sexism in society is getting

worse every year," and (h) "Society is set up so that men and women usually get what they deserve." Participants were asked to indicate the strength of agreement or disagreement with each of these items on a 9-point scale. Responses were coded in such a way that agreement with Items a, b, d, e, f, and h and disagreement with Items c and g resulted in higher scores on gender-specific system justification. The overall index was α = .65 in Jost & Kay' study (2005) and α= .85 in Calogero's study. Responses were made using scales from 1 (strongly disagree) to 9 (strongly agree). Items were averaged to create system-justification scores (α = .85). Scores ranged from 1 to 9, with higher scores indicating greater justification of the gender status quo. c. A set of feminist-activism items derived from prior research will be used to measure gender-based social activism (Stake, Roades, Rose, Ellis, & West, 1994). Participants rated eight items assessing the extent to which they had participated in various types of actions considered as social activism in the area of gender equality during the previous 6 months; responses were made on scales from 1 (never) to 7 (all the time). The eight items covered the following types of actions: discussing issues related to gender equality with friends or colleagues (in person or online—e.g., through e-mail, Facebook, Twitter, or MySpace); attending meetings, conferences, or workshops on gender-equality issues; signing a petition (in person or online) in support of women's rights and gender equality; circulating a petition (in person or online) related to women's rights or gender equality; handing out fliers related to women's-rights issues or gender equality; attending demonstrations, protests, or rallies related to women's rights or gender equality; working for women's rights campaigns (e.g., fund-raising); and acting as a spokesperson for a particular gender-equality issue. Ratings for all items were averaged to create social-activism scores (α = .92). Scores ranged from 1 to 7, with higher scores indicating more participation in gender-based social activism.

the statistical technique;

We will test and report simple correlation between our three variables of interest (i.e., self-objectification, system justification belief, collective action intentions) We will test and report a simple mediation model using PROCESS (Model 4; Hayes, 2013) to examine the direct and indirect effect of self-objectification on collective action intentions through system justification belief. Self-objectification will be entered as the predictor (X), collective action intentions will be entered as the criterion (Y), and system justification belief (M) will be entered as the mediating variable. Significance of indirect paths was assessed using 95% bias-corrected and accelerated confidence intervals with 10000 bootstrap resamples. We will also conduct and report Bayesian statistics to decide which of two hypotheses (H0 and H1) is more likely given data of PDR2 (we will test and report BF01) following the recommendation of Nera, Pantazi & Klein, (2018).

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

IV = Self-objectification Mediator = Gender-specific system justification DV = collective action intentions

rationale for each covariate used, if any;

Calogero does not mention any exclusion criteria in the original study

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

We used Bayesian statistics and based our conclusion on Nera, Pantazi & Klein, (2018). A BF01 greater than 3 can be consider as "some evidence", and BF01 greater than 10 as "strong evidence" in favor of H0.

Second Prediction

Describe the analyses that will test the second main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Third Prediction

Describe the analyses that will test the third main prediction from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Fourth Prediction

Describe the analyses that will test the fourth main prediction from the hypotheses

section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Further Predictions

Describe the analyses that will test any further (main) predictions from the hypotheses section. Include:

the relevant variables and how they are calculated;

No response

the statistical technique;

No response

each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate);

No response

rationale for each covariate used, if any;

No response

if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs toward making an evidential conclusion, including prior values or distributions.

No response

Recommended elements

Recommended Elements

Specify contingencies and assumptions, such as:

Method of correction for multiple tests.

N/A

The method of missing data handling (e.g., pairwise or listwise deletion, imputation, interpolation).

We will not consider participants with missing data in our analyses

Reliability criteria for item inclusion in scale.

Cronbach's Alpha inferior to .70 will lead to the exclusion of the scale

Anticipated data transformations.

Mean scores for continuous scales will be computed and the classical operation to calculate the "self-objectification" score will be run: Syntax for SPSS is: sum (NO1, NO2, NO3, NO4, NO5) – sum (O1, O2, O3, O4, O5) = sum (rank on non-observable items) – sum (rank on observable items). Scores on SOQ range from -25 (low score of self-objectification) to 25 (high score of self-objectification). Further, the transformation into Z scores will be necessary for the median analysis.

Assumptions of analyses, and plans for alternative/corrected analyses if each assumption is violated.

in case of non-normality of the data after excluding outliers, we will apply traditional mathematical operations to normalize our variables (e.g., log, square root)

Optionally, upload any files here that are related to your analyses (e.g., syntaxes, scripts, etc.).

No files selected

Final questions

Has data collection begun for this project?

No, data collection has not begun

If data collection has begun, have you looked at the data?

No

The (estimated) start and end dates for this project are

10/05/2019-20/05/2019

Any additional comments before I pre-register this project

No response