Il segnale e il rumore: Replicabilità dei risultati e potenza statistica

Marco Perugini

Roma, 27/9/2021

VERO TITOLO

Come indovinarci: Perché dovresti riflettere due volte prima di pianificare il tuo prossimo studio

Marco Perugini

Roma, 27/9/2021

Outline

- **Replicability in Psychology**
- **Sample planning**
- **Basic statistical concepts and errors of inference**
- **Power analysis**
  -Solving by sample or by effect size
  -Two independent groups and two repeated measures
  -Contrasts
  -Regression, Moderation
  -ANOVA (Between, Within, Mixed)
  -How to increase power
  -Uncertainty of estimates
  -Pointers for more complex models and designs
- **(Some) Tips for getting it right**

The problem

- Assume that, as scientists, we all want to get it right
- What can we do to increase our chances?

a) Get it right ≠ I am right

b) Get it right ≠ Get it published

Replicability in Psychology

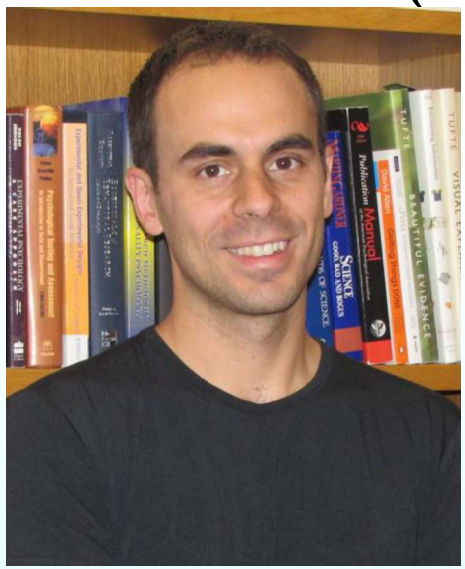
2011: A year to remember

- The year 2011 has been an *annus horribilis* for Psychology
- Three main events:

Bem

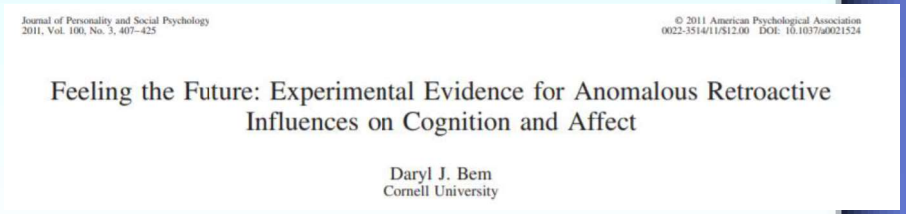
Stapel

Simmons (et al)



Bem (JPSP, 2011)

- Nine experiments showing ESP
- Strong reactions
- Initial article not replicating results was refused by JPSP
- Hard questions on the modal way of analyzing data and on "cherry-picking" results



Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

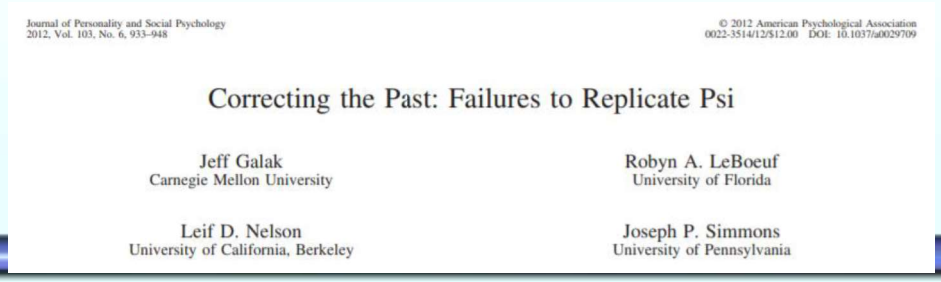
Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 426–432

0022-3514/11/$12.00 DOI: 10.1037/a0022790

Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas
University of Amsterdam

- Galak et al (2012): 7 failed replication attempts (n=3289)



Journal of Personality and Social Psychology
2012, Vol. 103, No. 6, 933–948

0022-3514/12/$12.00 DOI: 10.1037/a0029709

Correcting the Past: Failures to Replicate Psi

Jeff Galak
Carnegie Mellon University

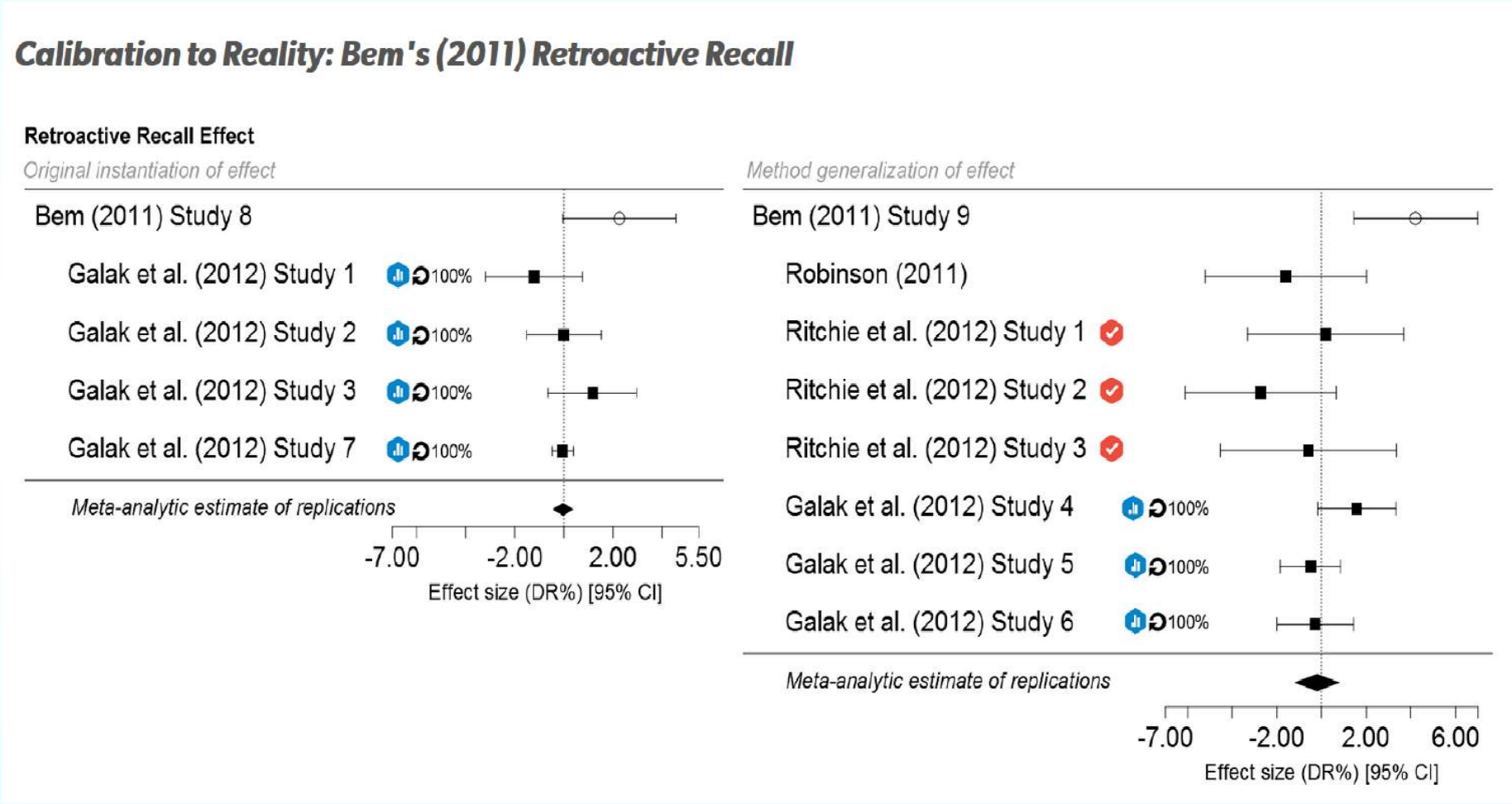
Robyn A. LeBoeuf
University of Florida

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

In the retroactive facilitation of recall studies, on the other hand, people are simply shown a list of words and are then asked to freely recall as many as possible. Participants are then randomly assigned to practice half of the words, with precognition being observed if people recall more of the words that they subsequently practice than words that they subsequently do not practice. In

Bem (2011)


Calibration to Reality: Bem's (2011) Retroactive Recall
Retroactive Recall Effect
Original instantiation of effect
Bem (2011) Study 8
Galak et al. (2012) Study 1
Galak et al. (2012) Study 2
Galak et al. (2012) Study 3
Galak et al. (2012) Study 7
100%
Meta-analytic estimate of replications
-7.00
-2.00
2.00
5.50
Effect size (DR%) [95% CI]
Method generalization of effect
Bem (2011) Study 9
Robinson (2011)
Ritchie et al. (2012) Study 1
Ritchie et al. (2012) Study 2
Ritchie et al. (2012) Study 3
Galak et al. (2012) Study 4
Galak et al. (2012) Study 5
Galak et al. (2012) Study 6
Meta-analytic estimate of replications
-7.00
-2.00
2.00
6.00
Effect size (DR%) [95% CI]

Stapel (September 2011)

- Resigned from Dean at Tilburg University (NL)
- Faked data: 53 retracted papers and 10 PhD thesis with invented or dubious data
- Levelt report (2012): proofs beyond doubts of faked data and strong criticisms to the scientific community
- Huge media impact

Simmons et al (PS, 2011)

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
22(11) 1359–1366

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

- They show that commonly used questionable research practices can allow to provide empirical evidence even for null effects (***false positive***)

Simmons et al (PS, 2011)

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level $p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Table 2. Simple Solution to the Problem of False-Positive Publications

Requirements for authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Guidelines for reviewers

1. Reviewers should ensure that authors follow the requirements.
2. Reviewers should be more tolerant of imperfections in results.
3. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
4. If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.

Huge scientific impact (3130 citations at 31/08/21, one of the most cited papers from 2011 in all Psychology). Not all solutions are equally convincing, but they make many good points

The replicability crisis

- From 2011 increasing appreciation of problems in published research in top journals in Psychology
- Cases of not replicated results and outright frauds
- Fraud is a problem, but not the only one
- Hard questions on the modal way of analyzing data and on “cherry-picking” results
- Important advances in research methodology
- Rapid changes in standards for research and for publishing

Replicability

- If a result is not replicated, it is not valid
- To be replicated, it needs to be replicable

Replicability

- A key concept in Science
- Almost forgotten in Psychology
- Now at the forefront
- What is replicability?

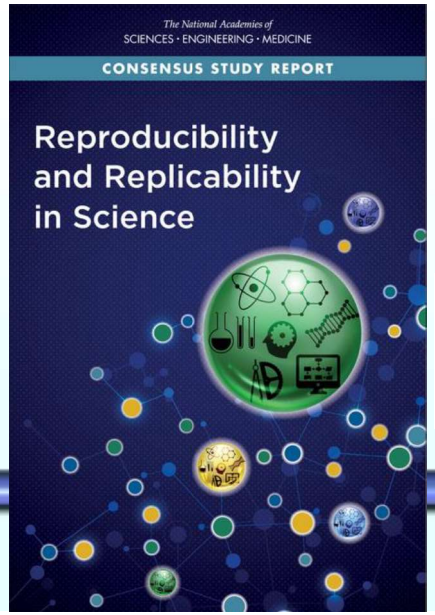
Replicability

European Journal of Personality, *Eur. J. Pers.* **27**: 108–119 (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.1919

Recommendations for Increasing Replicability in Psychology†

JENS B. ASENDORPF[1]*, MARK CONNER[2], FILIP DE FRUYT[3], JAN DE HOUWER[4], JAAP J. A. DENISSEN[5], KLAUS FIEDLER[6], SUSANN FIEDLER[7], DAVID C. FUNDER[8], REINHOLD KLIEGL[9], BRIAN A. NOSEK[10], MARCO PERUGINI[11], BRENT W. ROBERTS[12], MANFRED SCHMITT[13], MARCEL A. G. VANAKEN[14], HANNELORE WEBER[15] and JELTE M. WICHERTS[5]

[1]*Department of Psychology, Humboldt University Berlin, Berlin, Germany*
[2]*Institute of Psychological Sciences, University of Leeds, Leeds, UK*
[3]*Department of Developmental, Personality and Social Psychology, Ghent University, Ghent, Belgium*
[4]*Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium*
[5]*School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands*
[6]*Department of Psychology, University of Heidelberg, Heidelberg, Germany*
[7]*Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[8]*Department of Psychology, University of California at Riverside, Riverside, CA USA*
[9]*Department of Psychology, University of Potsdam, Potsdam, Germany*
[10]*Department of Psychology, University of Virginia, Charlottesville, VA USA*
[11]*Department of Psychology, University of Milano-Bicocca, Milan, Italy*
[12]*Department of Psychology, University of Illinois, Chicago, IL USA*
[13]*Department of Psychology, University of Koblenz–Landau, Landau, Germany*
[14]*Department of Psychology, Utrecht University, Utrecht, The Netherlands*
[15]*Department of Psychology, University of Greifswald, Greifswald, Germany*



The National Academies of | SCIENCES ENGINEERING MEDICINE

Reproducibility and Replicability in Science
(2019)
Consensus Study Report

GIORNALE ITALIANO DI PSICOLOGIA / a. XLI, n. 1, marzo 2014

LA CRISI INTERNAZIONALE DI CREDIBILITÀ DELLA PSICOLOGIA COME UN'OPPORTUNITÀ DI CRESCITA: PROBLEMI E POSSIBILI SOLUZIONI

MARCO PERUGINI

Università Milano-Bicocca

GIORNALE ITALIANO DI PSICOLOGIA / a. XLV, n. 4, dicembre 2018

Definitions

- **Reproducibility**: Obtaining consistent computational results using the same input data, computational steps, methods, and code, and conditions of analysis

- **Replicability**: obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data

Conditions for Replicability

- The study should be described in a way such that everyone qualified can replicate it
- This implies a very detailed method section, with information that often is not disclosed
- Also, the data should be publicly available (at a minimum upon request) to replicate the results using appropriate analyses (*reproducibility*)


Behav Res (2016) 48:1205–1226
DOI 10.3758/s13428-015-0664-2

The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nuijten[1] · Chris H. J. Hartgerink[1] · Marcel A. L. M. van Assen[1] · Sacha Epskamp[2] · Jelte M. Wicherts[1]

- Transparency in research

Replicability as a pre-condition for validity

- If an effect is not replicable, it cannot be valid (according to scientific standards)
- If an effect is replicable, it may or may not be valid
- Validity assumes but goes beyond replicability
- Analogy with psychometric measures: Reliability is a necessary but insufficient condition for validity of a measure

Replicability: a continuum

Generalizibility
Validity

LEBEL, BERGER, CAMPBELL, AND LOVING

Highly similar ← *Replication continuum* → *Highly dissimilar*

	Direct replication			Conceptual replication	
Design facet	**Exact replication** (Everything controllable the same)	**Very close replication** (Procedure or context is different)	**Close replication** (IV or DV stimuli are different)	**Far replication** (IV or DV operationalization is different)	**Very Far replication** (Everything can be different)
IV operationalization	same	same	same	different	
DV operationalization	same	same	same	different	
IV stimuli	same	same	different		
DV stimuli	same	same	different		
Procedural details	same	different			
Physical setting	same	different			
Contextual variables	different				
⋮	⋮				

Figure 1. A simplified replication taxonomy to guide the classification of relative methodological similarity of a replication study to an original study. "Same" ("different") indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. "Everything controllable" indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.). See the online article for the color version of this figure.

The Reproducibility Project

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

- Open Science Center
- 270 volunteers, 64 universities, 11 countries, 100 replicated studies from 3 main journals (JPSP, PS, JEP:LMC) in 2008.
- Started in 2011, published in 2015 in *Science*
- **What results one can expect?**

What counts as a replication?

1. Effect itself against null (p-values)
2. Effect size comparison
3. Meta-analytic precision estimate

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

Main results

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < 0.05$" (3 original nulls excluded; $n = 97$ studies); "mean original and replication effect sizes" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); "percent meta-analytic ($P < 0.05$)" (25 excluded; $n = 75$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

		Effect size comparison						Original and replication combined			
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Median original *df/N*	Mean (SD) replication effect size	Median replication *df/N*	Average replication power	Meta-analytic mean (SD) estimate	Percent meta-analytic ($P < 0.05$)	Percent original effect size within replication 95% CI	Percent subjective "yes" to "Did it replicate?"
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92	0.309 (0.223)	68	47	39
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91	0.138 (0.087)	43	34	25
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93	0.393 (0.209)	86	62	54
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92	0.286 (0.228)	58	40	32
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94	0.464 (0.221)	92	60	53

Effect size comparison

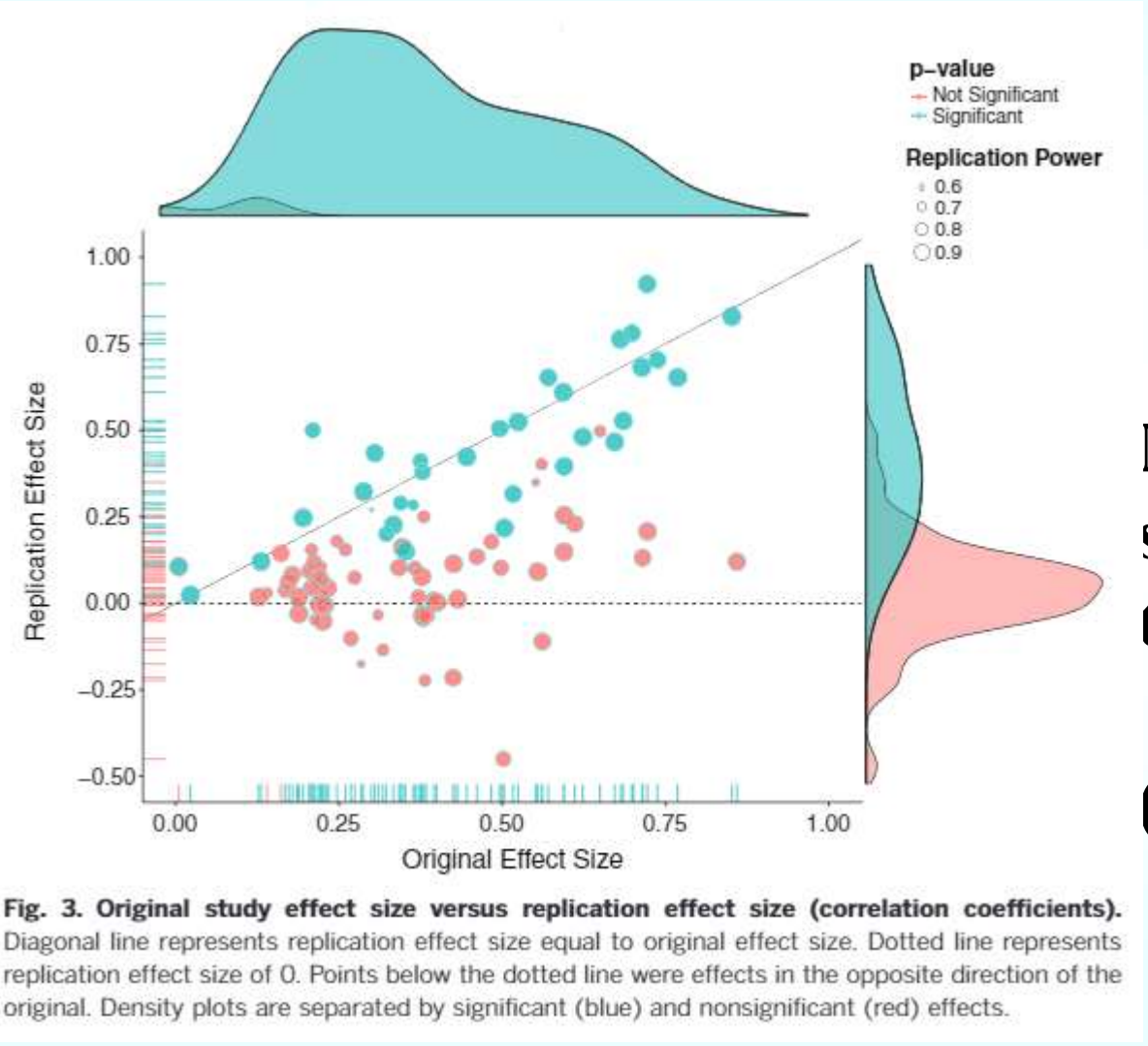


Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

r(spearman) = 0.51

Most replications smaller effect size than original

(d=0.20 vs. 0.40)

Summing up RP main results

- 36% replicate at p<.05 (simple answer)
- Effect size are half (publication bias, file drawer effect)
- Less likely to replicate if weaker evidence in original study (p<.05 worse than p<.001)
- Milestone achievement of Psychology
- Followed by RPs in other scientific domains and many other RPs in Psychology
  (over 1000 replication studies from 2011)

Is the problem unique to Psychology?

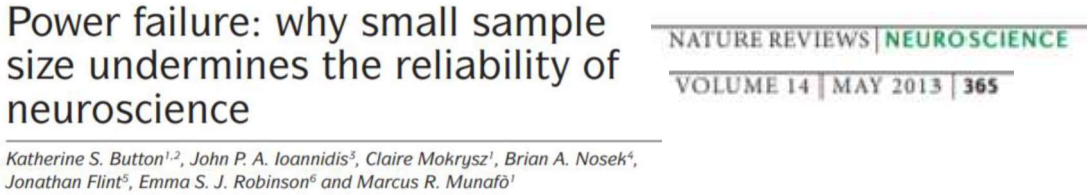
- **NO !!** (Ioannidis, 2005)


Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine | www.plosmedicine August 2005 | Volume 2 | Issue 8 | e124

- Average power in **Neuroscience**: .21 (Button et al., 2013)
  This means around 1/5 chance of positive findings (which means there must be many published false-positive findings…)


Power failure: why small sample size undermines the reliability of neuroscience

NATURE REVIEWS | NEUROSCIENCE

VOLUME 14 | MAY 2013 | 365

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

- **Cancer Biology**: Replication rate of main results from pre-clinical trails (Begley & Ellis, 2012): from 11% to 25%. Recent reproducibility study ongoing (?)


IN FOCUS NEWS

CANCER BIOLOGY

Reproducibility project yields muddy results

An ambitious effort to replicate cancer studies is provoking controversy.

BY MONYA BAKER AND ELIE DOLGIN


CRITIQUE & DEBATE

National Science Review
5: 619–624, 2018
doi: 10.1093/nsr/nwy021
Advance access publication 2 February 2018

MOLECULAR BIOLOGY & GENETICS

On the low reproducibility of cancer studies

Haijun Wen[1], Hurng-Yi Wang[2], Xionglei He[1] and Chung-I Wu[1,3,*]

Replicability Projects

RESEARCH ARTICLE

SCIENCE

PSYCHOLOGY

Estimating the reproducibi psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to wh current research is unknown. We conducted replications of 100 exp

Social Psychology 2014; Vol. 45(3):142–152
DOI: 10.1027/1864-9335/a000178

Replication

ECONOMICS

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

Review of Philosophy and Psychology
pp 1–36 | Cite as

Estimating the Reproducibility of Experimental Philosophy

Authors | Authors and affiliations

Florian Cova, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Be

Renatas Berniūnas, Jordane Boudesseul, Matteo Colombo, Fiery Cushman, Rodrigo Diaz, Noah N'Djaye

nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 5

reber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16],

Psychological Science
2019, Vol. 30(5) 711–727
© The Author(s) 2019
Article reuse guidelines:

Registered Replication Report

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project

TC

Christopher J. Soto
Department of Psychology, Colby College

Many Labs 2: Investigating Variation in Replicability Across Samples and Settings

Richard A. Klein[1], Michelangelo Vianello[2], Fred Hasselman[3,4],

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(4) 443–490
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245918810225
www.psychologicalscience.org/AMPPS

SAGE

RP Experimental Economics

ECONOMICS

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

The replicability of some scientific findings has recently been called into question. To contribute data about replicability in economics, we replicated 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014 All of these replications followed predefined analysis plans that were made publicly available beforehand, and they all have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original. The replicability rate varies between 67% and 78% for four additional replicability indicators, including a prediction market measure of peer beliefs.

SCIENCE sciencemag.org

25 MARCH 2016 • VOL 351 ISSUE 6280 1433

61% studies replicated (significant p value)

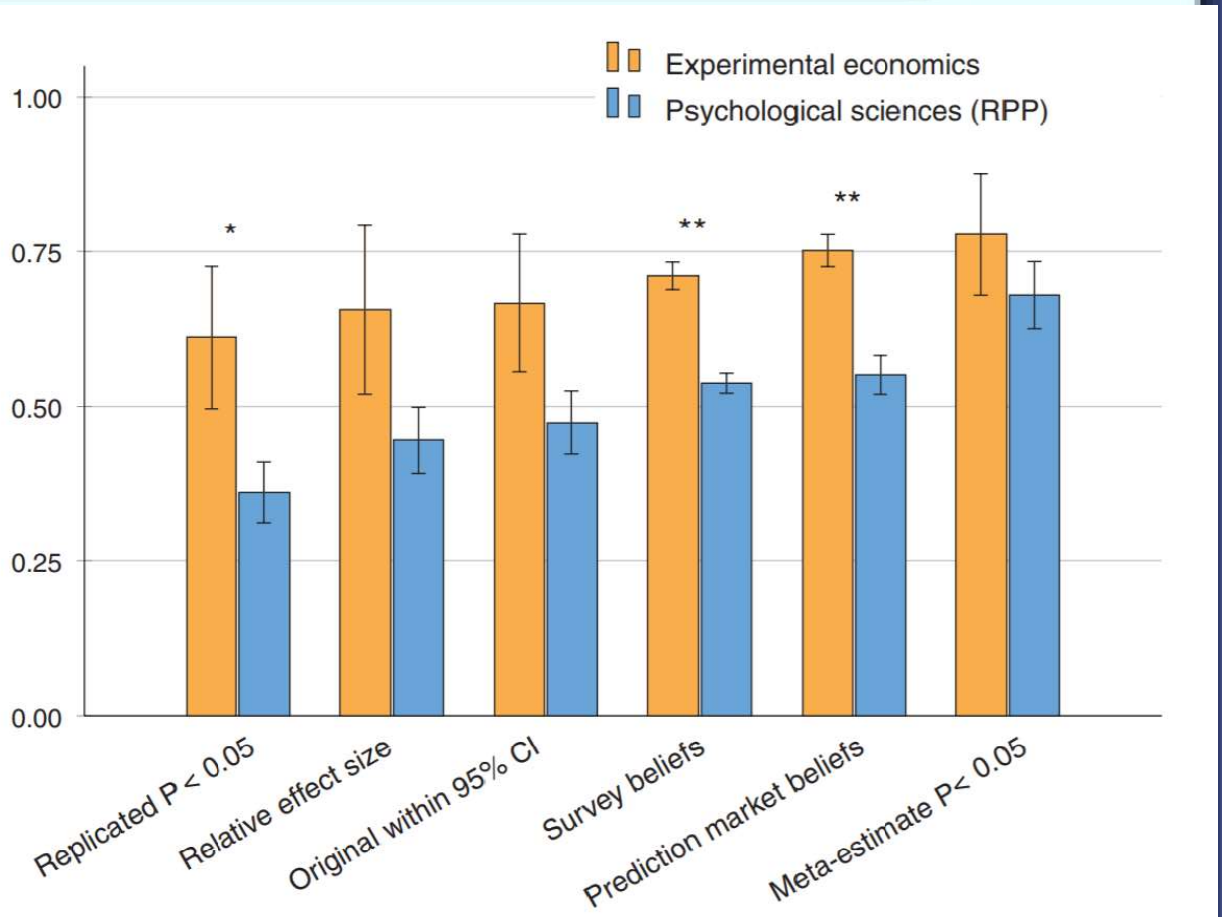
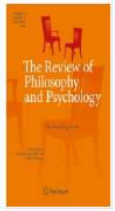


Fig. 4. A comparison of replicability indicators in experimental economics (this study) and psychological sciences (RPP). The graph shows means ± SE for replicability indicators. All six replicability indicators are higher for experimental economics; this difference is significant for three of the replicability indicators. The average difference in replicability across the six indicators is 19 percentage points. Details about the statistical tests are included in the supplementary materials. *$P < 0.05$; **$P < 0.01$.

RP Experimental Philosophy



Review of Philosophy and Psychology

pp 1–36 | Cite as

Estimating the Reproducibility of Experimental Philosophy

Authors | Authors and affiliations

Florian Cova, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Be

Renatas Berniūnas, Jordane Boudesseul, Matteo Colombo, Fiery Cushman, Rodrigo Diaz, Noah N'Djaye I

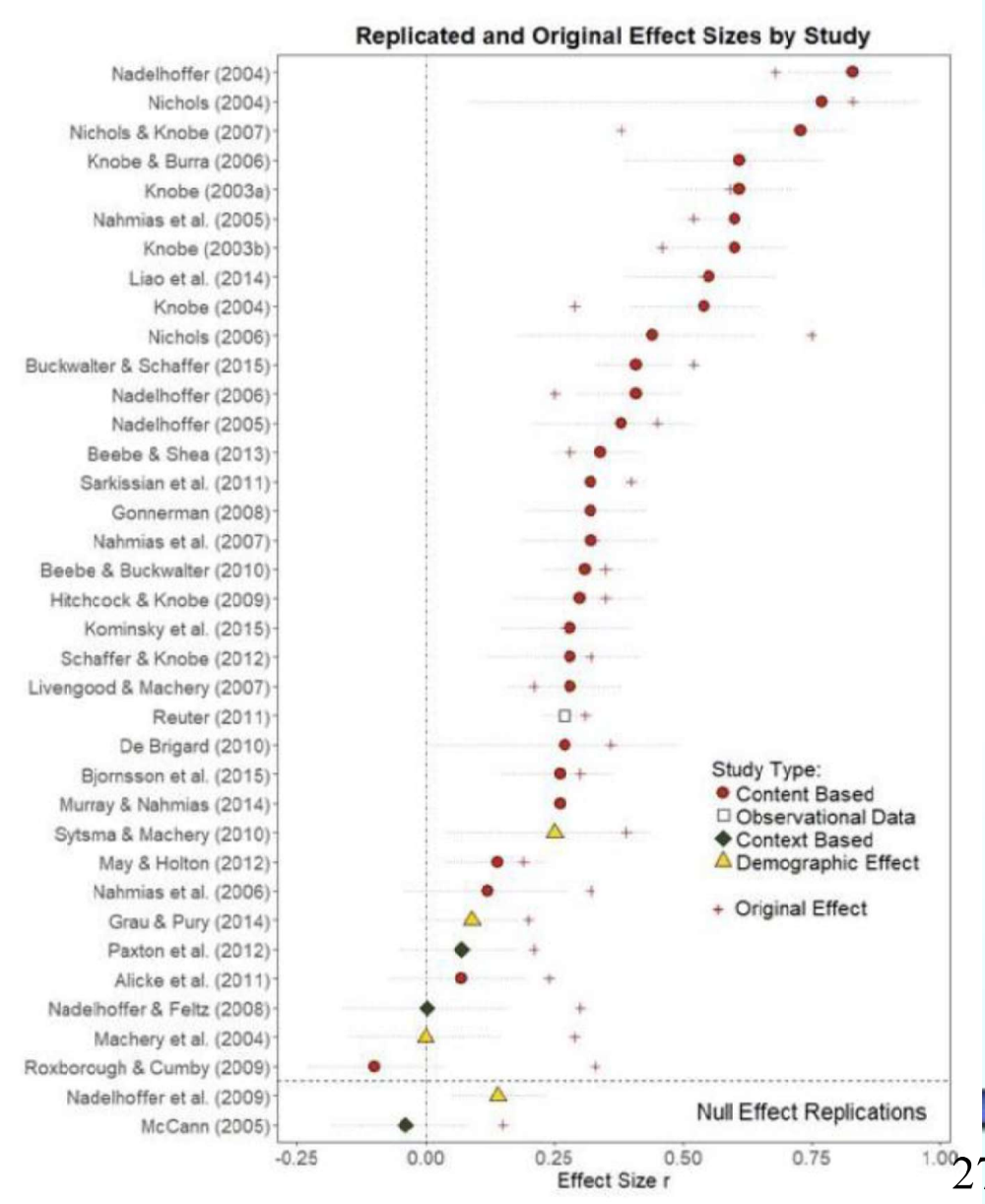
Vilius Dranseika, Brian D. Earp, show 27 more

Article

First Online: 14 June 2018

16 Shares | 450 Downloads | 1 Citations

75% studies replicated (significant p value)



RP Social Science in N & S

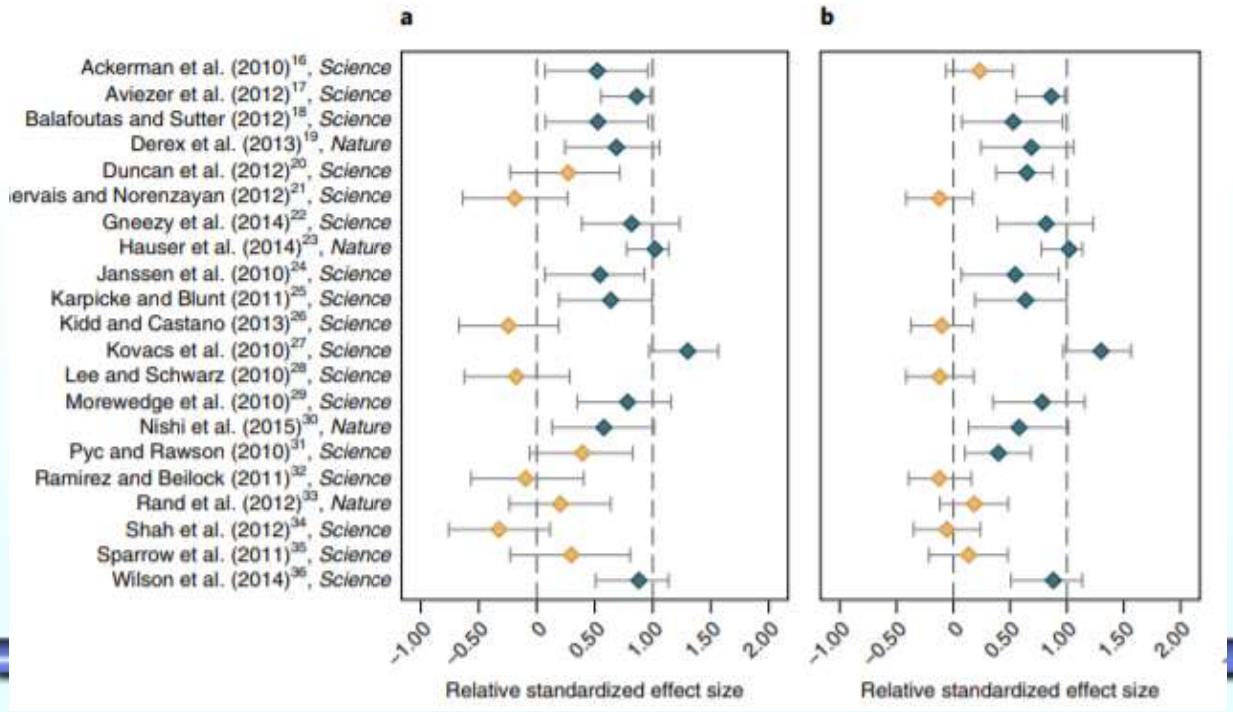
nature human behaviour

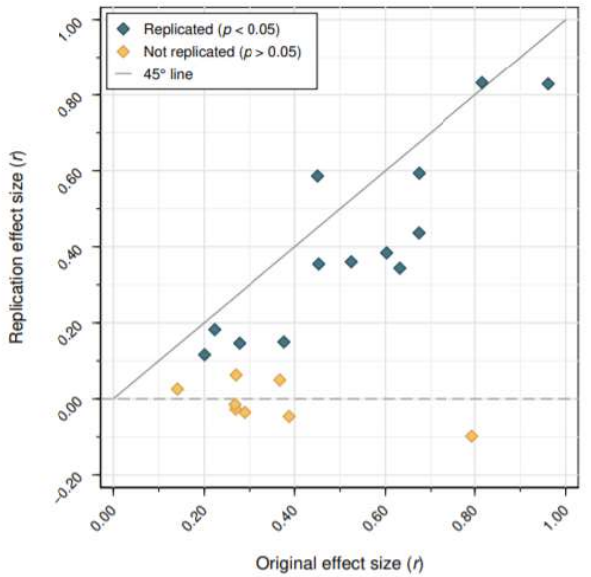
LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16]*, Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]





62% studies replicated **(significant p value)**

Effect Size is around half **(d=0.51 vs. 1.04)**

Ego-Depletion

A Multilab Preregistered Replication of the Ego-Depletion Effect

Perspectives on Psychological Science
2016, Vol. 11(4) 546–573

DOI: 10.1177/1745691616652873
pps.sagepub.com
SAGE

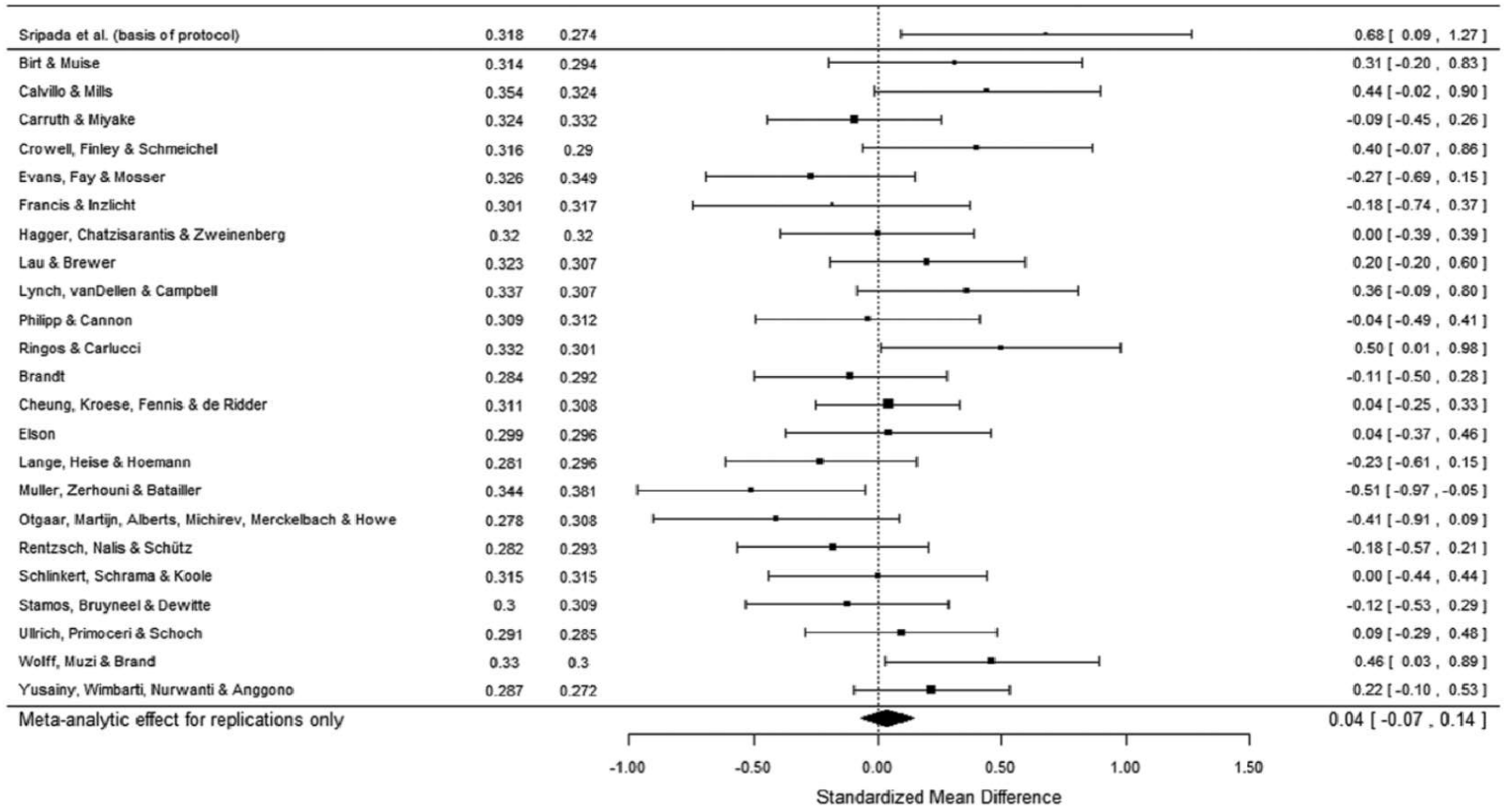
Martin S. Hagger and Nikos L. D. Chatzisarantis
Curtin University, Australia

Contributing authors: Hugo Alberts, Calvin Octavianus Anggono, Cédric Batailler, Angela R. Birt, Ralf Brand, Mark J. Brandt, Gene Brewer, Sabrina Bruyneel, Dustin P. Calvillo, W. Keith Campbell, Peter R. Cannon, Marianna Carlucci, Nicholas P. Carruth, Tracy Cheung, Adrienne Crowell, Denise T. D. De Ridder, Siegfried Dewitte, Malte Elson, Jacqueline R. Evans, Benjamin A. Fay, Bob M. Fennis, Anna Finley, Zoë Francis, Elke Heise, Henrik Hoemann, Michael Inzlicht, Sander L. Koole, Lina Koppel, Floor Kroese, Florian Lange, Kevin Lau, Bridget P. Lynch, Carolien Martijn, Harald Merckelbach, Nicole V. Mills, Alexej Michirev, Akira Miyake, Alexandra E. Mosser, Megan Muise, Dominique Muller, Milena Muzi, Dario Nalis, Ratri Nurwanti, Henry Otgaar, Michael C. Philipp, Pierpaolo Primoceri, Katrin Rentzsch, Lara Ringos, Caroline Schlinkert, Brandon J. Schmeichel, Sarah F. Schoch, Michel Schrama, Astrid Schütz, Angelos Stamos, Gustav Tinghög, Johannes Ullrich, Michelle vanDellen, Supra Wimbarti, Wanja Wolff, Cleoputri Yusainy, Oulmann Zerhouni, Maria Zwienenberg

Meta-analysis (Hagger et al., 2010) did already show a significant effect for ego-depletion (*d=0.62*)

Yet, meta-analysis is no substitute for replication

Ego-Depletion



Ego-Depletion again

A Multi-Site Preregistered Paradigmatic Test of the Ego Depletion Effect

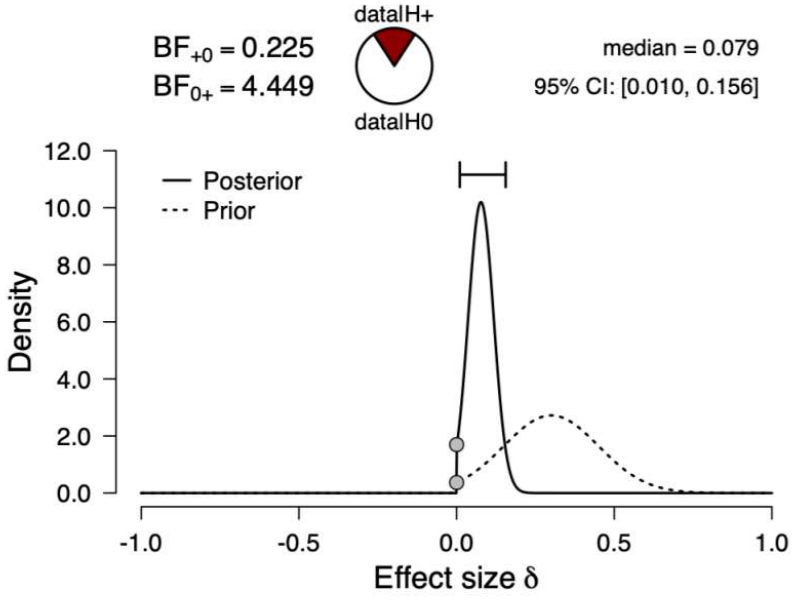
in press, *Psychological Science*

Vohs, Kathleen D., University of Minnesota
Schmeichel, Brandon J., Texas A&M University
Lohmann, Sophie, Max Planck Institute for Demographic Research and University of Illinois at Urbana-Champaign
Gronau, Quentin, F., University of Amsterdam
Finley, Anna, J. University of Wisconsin-Madison

Paradigmatic means using the **best** conceptual manipulation.
No evidence or perhaps some evidence for very small effect

Table 4. *Depletion Effect: Frequentist Meta-Analyses*

		Random-effects meta-analysis			Fixed-effects meta-analysis	
DV	*N*	d	CI	I^2%	d	CI
Overall depletion effect	2461	0.06	[-0.02, 0.14]	2.54	0.06	[-0.02, 0.14]
Overall figure tracing performance	1216	0.12	[-0.01, 0.24]	15.16	0.11	[-0.00, 0.22]



Meta-analysis vs. Replication

nature human behaviour

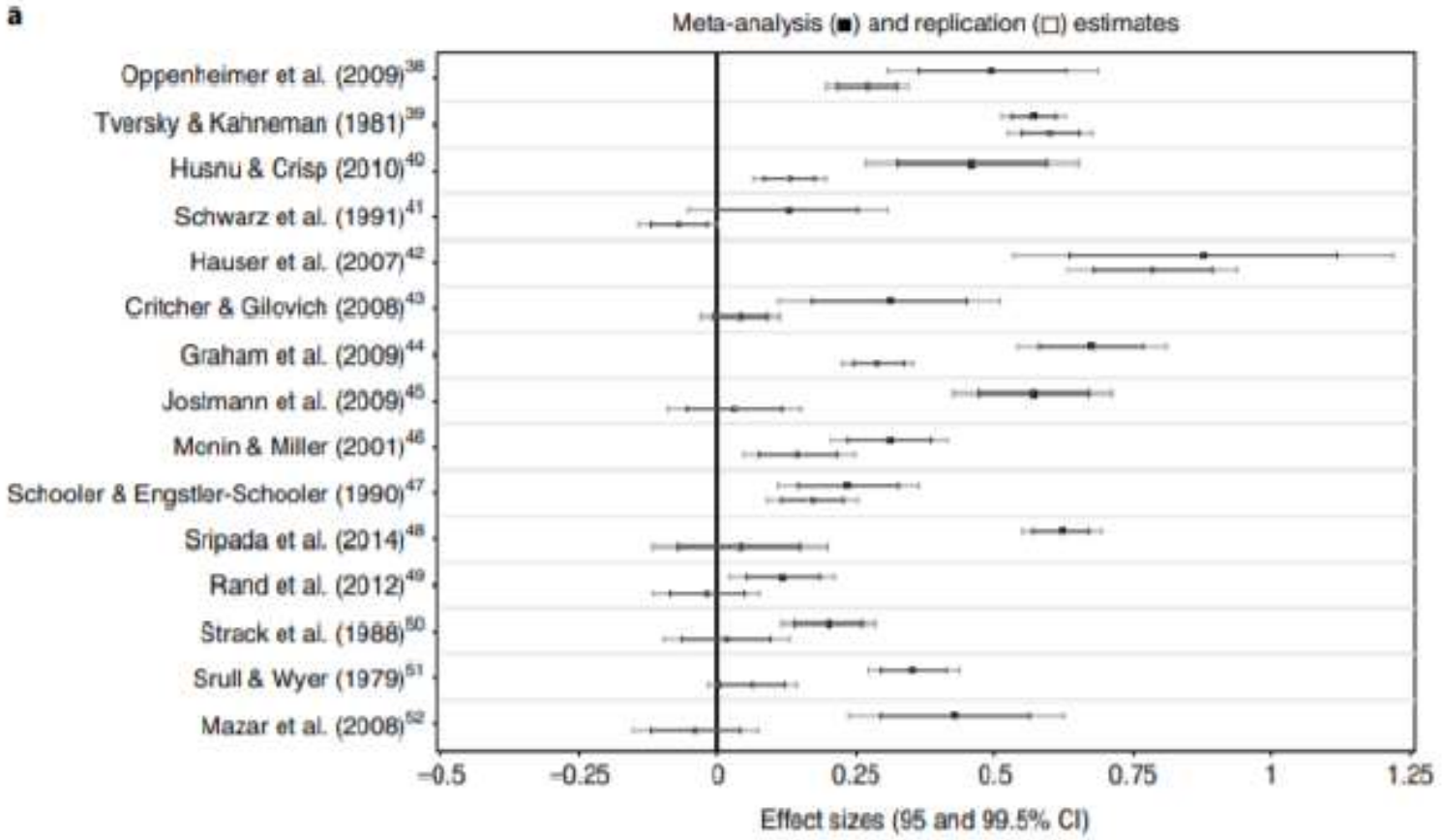
ARTICLES

https://doi.org/10.1038/s41562-019-0787-z

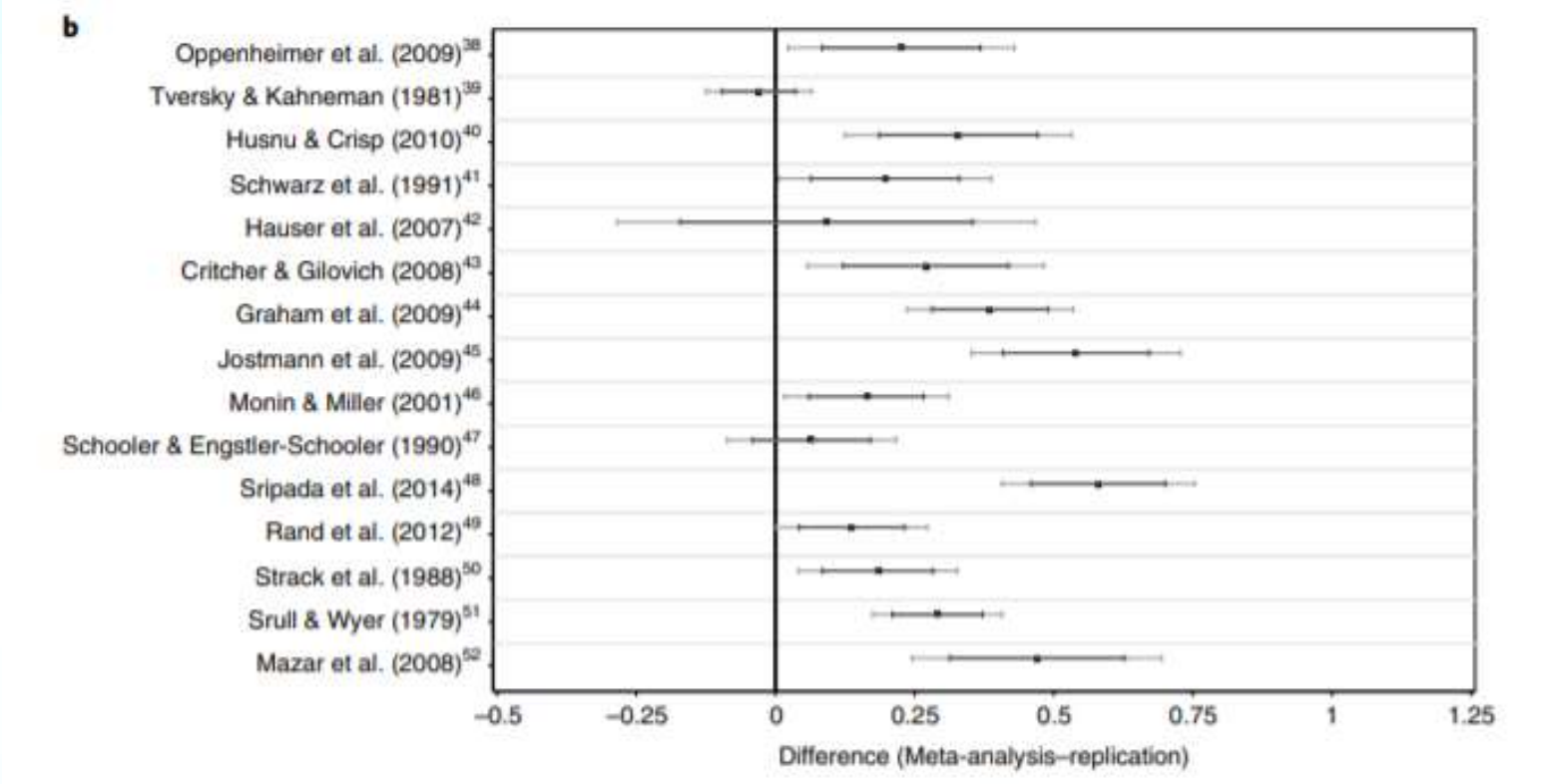
Corrected: Author correction

Comparing meta-analyses and preregistered multiple-laboratory replication projects

Amanda Kvarven[1,3], Eirik Strømland[1,3] and Magnus Johannesson[2*]



Meta-analysis vs. Replication



7/15 (47%) replicated significant effect

12/15 (80%) smaller ES

Average ES (Cohen's d): 0.16 vs. 0.42

Money Priming ! (Vohs et al)

Journal of Experimental Psychology: General
2013, Vol. 142, No. 2, 301–306

0096-3445/13/$12.00 DOI: 10.1037/a0029288

BRIEF REPORT

Mere Exposure to Money Increases Endorsement of Free-Market Systems and Social Inequality

Eugene M. Caruso
University of Chicago

Kathleen D. Vohs
University of Minnesota

Brittani Baxter
University of Chicago

Adam Waytz
Northwestern University

REPORTS 17 NOVEMBER 2006 VOL 314 SCIENCE www.sciencemag.org

The Psychological Consequences of Money

Kathleen D. Vohs,[1]* Nicole L. Mead,[2] Miranda R. Goode[3]

Money has been said to change people's motivation (mainly for the better) and their behavior toward others (mainly for the worse). The results of nine experiments suggest that money brings about a self-sufficient orientation in which people prefer to be free of dependency and dependents. Reminders of money, relative to nonmoney reminders, led to reduced requests for help and reduced helpfulness toward others. Relative to participants primed with neutral concepts, participants primed with money preferred to play alone, work alone, and put more physical distance between themselves and a new acquaintance

CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE

Merely Activating the Concept of Money Changes Personal and Interpersonal Behavior

Kathleen D. Vohs,[1] Nicole L. Mead,[2] and Miranda R. Goode[3]

[1]Department of Marketing, Carlson School of Management, University of Minnesota, [2]Department of Psychology, Florida State University, and [3]Sauder School of Business, University of British Columbia

Money Priming example

Journal of Experimental Psychology: General
2013, Vol. 142, No. 2, 301–306

0096-3445/13/$12.00 DOI: 10.1037/a0029288

BRIEF REPORT

Mere Exposure to Money Increases Endorsement of Free-Market Systems and Social Inequality

Eugene M. Caruso
University of Chicago

Kathleen D. Vohs
University of Minnesota

Brittani Baxter
University of Chicago

Adam Waytz
Northwestern University

Background-image conditions. Participants assigned to the *background-image/money condition* saw a faint image of $100 bills in the background of the initial instruction screen, whereas participants assigned to the *background-image/control condition* saw a blurred version of this image, such that the bills were unrecognizable (Caruso et al., 2013; for a similar manipulation, see Kushlev, Dunn, & Ashton-James, 2012).

Money Prime

Control Prime

Figure 1. Images used for the money prime condition and control condition (Experiments 1, 4, and 5).

Money Priming ???

Show Me the Money: A Systematic Exploration of Manipulations, Moderators, and Mechanisms of Priming Effects

Psychological Science
2017, Vol. 28(8) 1148–1159

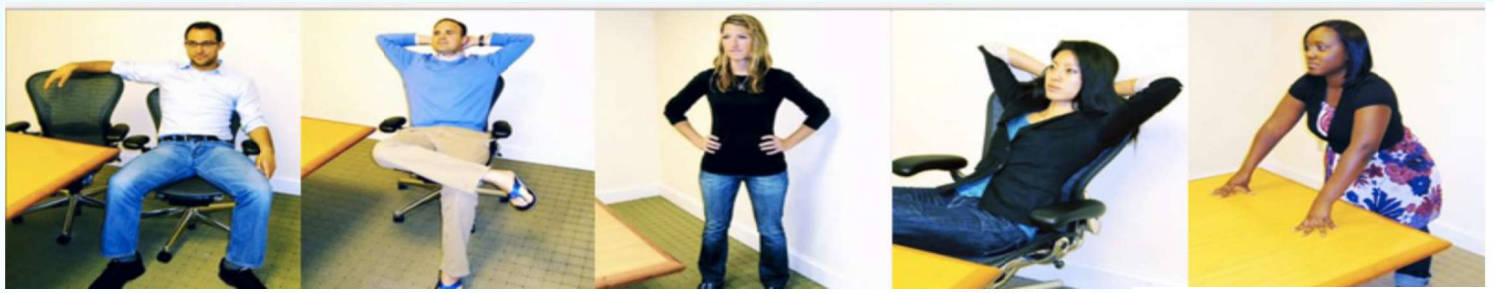
DOI: 10.1177/0956797617706161
www.psychologicalscience.org/PS
SAGE

Eugene M. Caruso[1], Oren Shapira[2], and Justin F. Landy[1]
[1]Booth School of Business, University of Chicago, and [2]Department of Medicine, Stony Brook University

Abstract

A major challenge for accumulating knowledge in psychology is the variation in methods and participant populations across studies in a single domain. We offer a systematic approach to addressing this challenge and implement it in the domain of money priming. In three preregistered experiments (N = 4,649), participants were exposed to one of a number of money manipulations before completing self-report measures of money activation (Study 1); engaging in a behavioral-persistence task (Study 3); completing self-report measures of subjective wealth, self-sufficiency, and communion-agency (Studies 1–3); and completing demographic questions (Studies 1–3). Four of the five manipulations we tested activated the concept of money, but, contrary to what we expected based on the preponderance of the published literature, no manipulation consistently affected any dependent measure. Moderation by sociodemographic characteristics was sparse and inconsistent across studies. We discuss implications for theories of money priming and explain how our approach can complement recent efforts to build a reproducible, cumulative psychological science.

Power posing (Amy Cuddy)



Research Report

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

Dana R. Carney[1], Amy J.C. Cuddy[2], and Andy J. Yap[1]
[1]Columbia University and [2]Harvard University

Psychological Science
21(10) 1363–1368
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797610383437
http://pss.sagepub.com

Method

Participants and overview of procedure

Forty-two participants (26 females and 16 males) were randomly assigned to the high-power-pose or low-power-pose condition. Participants believed that the study was about the

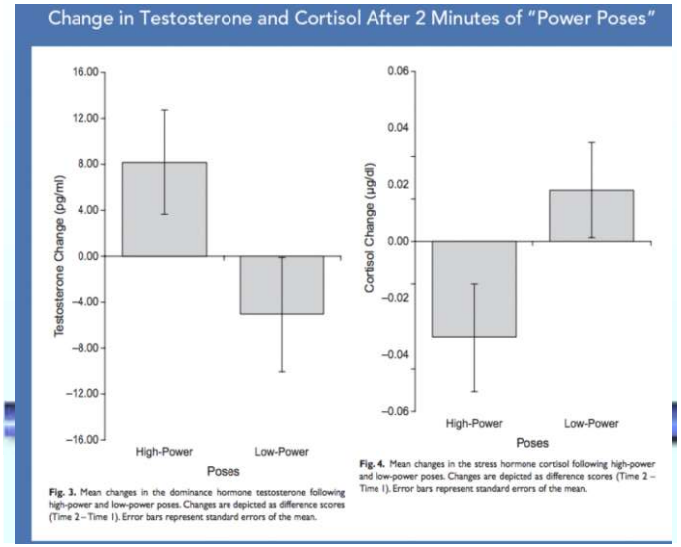


Fig. 3. Mean changes in the dominance hormone testosterone following high-power and low-power poses. Changes are depicted as difference scores (Time 2 – Time 1). Error bars represent standard errors of the mean.

Fig. 4. Mean changes in the stress hormone cortisol following high-power and low-power poses. Changes are depicted as difference scores (Time 2 – Time 1). Error bars represent standard errors of the mean.

Commentary

Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women

Eva Ranehill[1], Anna Dreber[2], Magnus Johannesson[2], Susanne Leiberg[1], Sunhae Sul[3], and Roberto A. Weber[1]
[1]Department of Economics, University of Zurich; [2]Department of Economics, Stockholm School of

Psychological Science
2015, Vol. 26(5) 653–656
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797614553946
pss.sagepub.com

sample size of 100 participants would be suitable. On the basis of the results of these first 100 observations, we decided to collect data from another 100 participants to further increase the reliability of our results. Of the final sample of 200, 98 were women and 102 men.

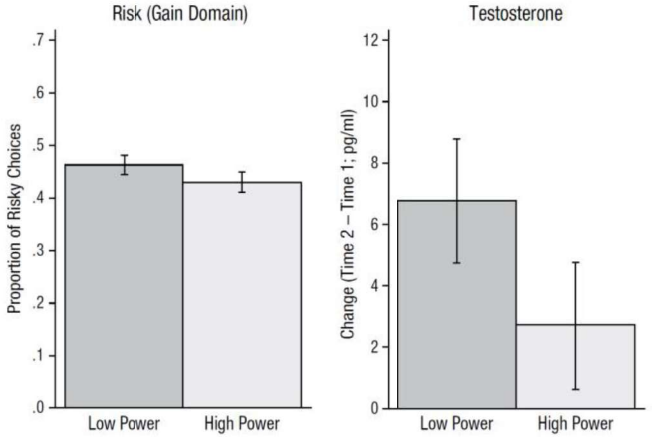


Fig. 1. Mean proportion of risky choices in the gain domain (left) and mean change in testosterone from before the power-pose manipulation (Time 1) to 17 min after the power-pose manipulation (Time 2; right). For each graph, results are shown separately for the high- and low-power-pose conditions. Error bars represent standard errors of the mean.

COMPREHENSIVE RESULTS IN SOCIAL PSYCHOLOGY, 2017
VOL. 2, NO. 1, 1–5
https://doi.org/10.1080/23743603.2017.1309876

INTRODUCTION

CRSP special issue on power poses: what was the point and what did we learn?

Joseph Cesario[a], Kai J. Jonas[b] and Dana R. Carney[c]
[a]Department of Psychology, Michigan State University, East Lansing, USA; [b]Work and Social Psychology Department, Maastricht University, Maastricht, Netherlands; [c]Haas School of Business, University of California, Berkeley, USA

- Bailey, LaFrance, and Dovidio (2017) sought to investigate an interaction of power posing, target gender, and participant gender. They did not replicate the effect of power poses on risky behavior.
- Bombari, Schmid Mast, and Pulfrey (2017) planned to test whether imagined or performed power poses had similar effects. They did not replicate the effect of power poses on risky behavior.
- Klaschinski, Schnabel, and Schröder-Abé (2017) wanted to replicate the effects of power posing on dominance and social sensitivity in an interview context, but they did not replicate the effects.
- Jackson, Nault, Smart Richman, LaBelle, and Rohleder (2017) sought to test the effect of power posing on self-concept. Although a preliminary study obtained an interesting effect, they did not replicate this in the higher-powered, preregistered study.
- Keller, Johnson, and Harder (2017) wanted to test whether awareness of the function of power poses moderates their effectiveness. They did not replicate the basic power pose effect.
- Latu, Duffy, Pardal, and Alger (2017) tested an interesting dependent variable in the context of power poses, persuasive messages. They did not observe any effect of power poses on persuasive message perception.
- Ronay, Tybur, van Huijstee, and Morssinkhoff (2017) wanted to investigate the mediating role of testosterone and overconfidence on the link between power posing and risk taking, but they did not replicate the effect.

Power posing (Dana Carney)

Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance DR Carney, AJC Cuddy, AJ Yap Psychological science 21 (10), 1363-1368	1356	2010

My position on "Power Poses"

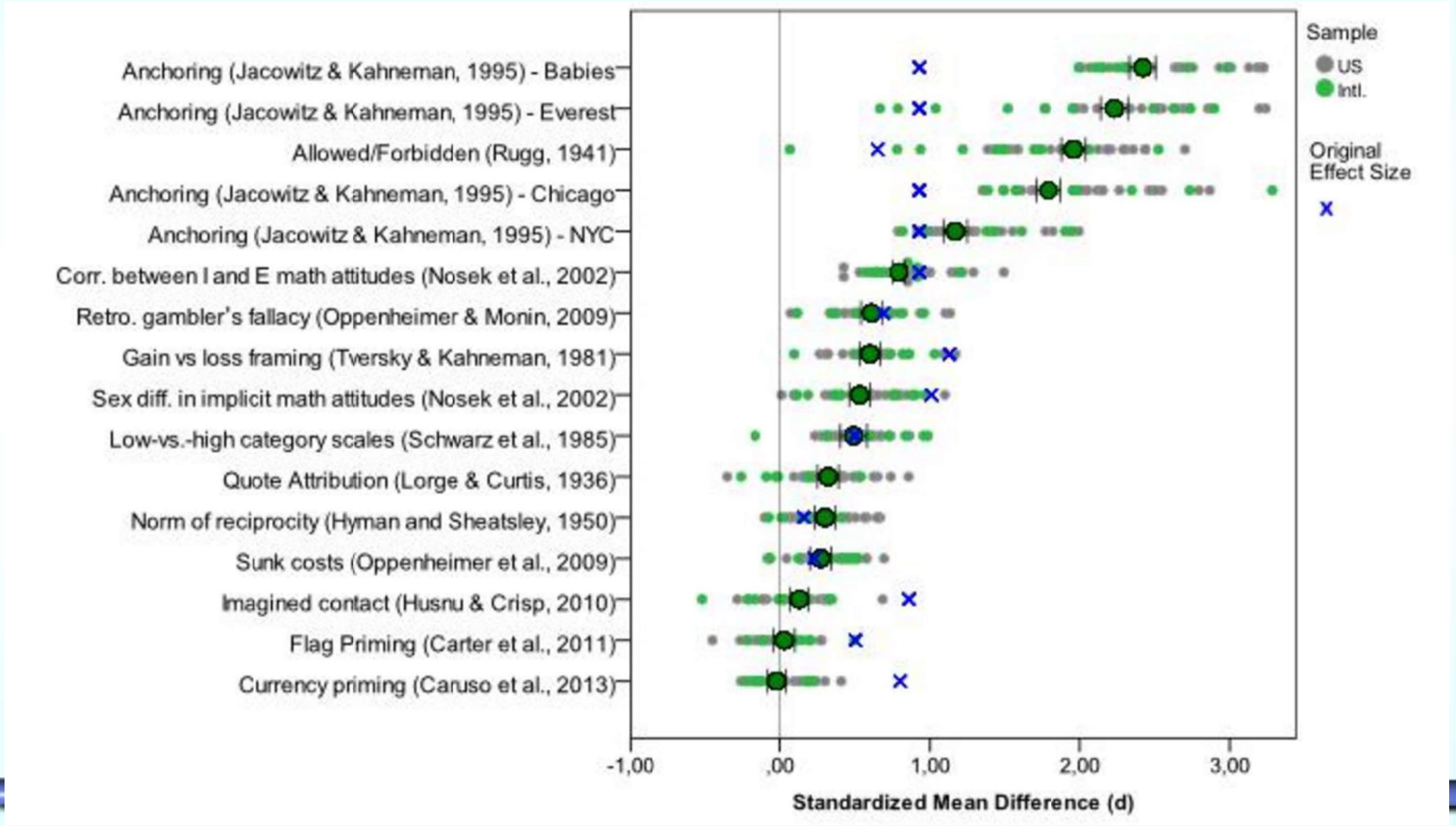
Regarding: Carney, Cuddy & Yap (2010).

Reasonable people, whom I respect, may disagree. However since early 2015 the evidence has been mounting suggesting there is unlikely any embodied effect of nonverbal expansiveness (vs. contractiveness)—i.e.., "power poses" -- on internal or psychological outcomes.

As evidence has come in over these past 2+ years, my views have updated to reflect the evidence. As such, **I do not believe that "power pose" effects are real.**

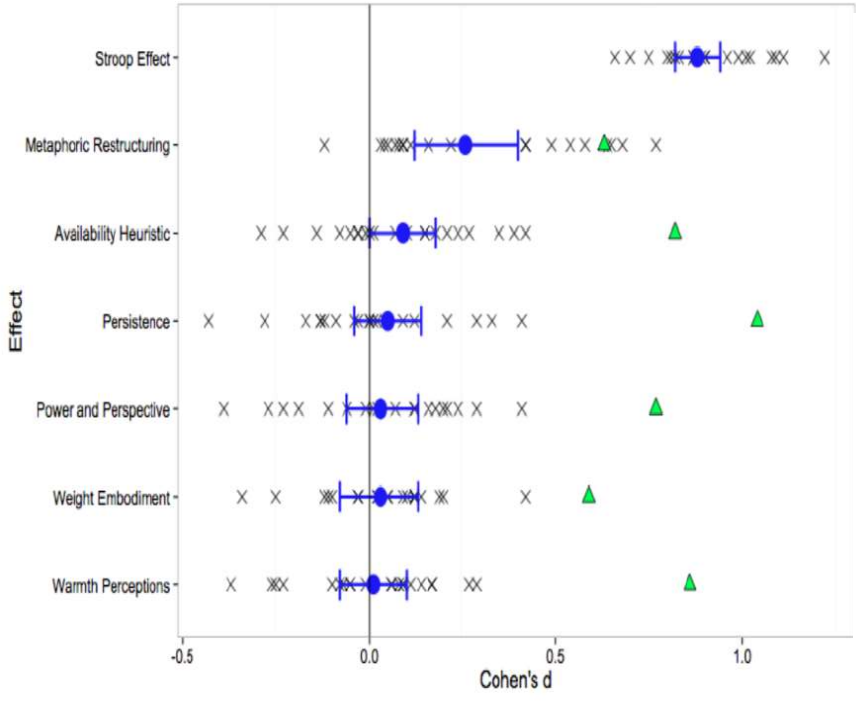
https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf

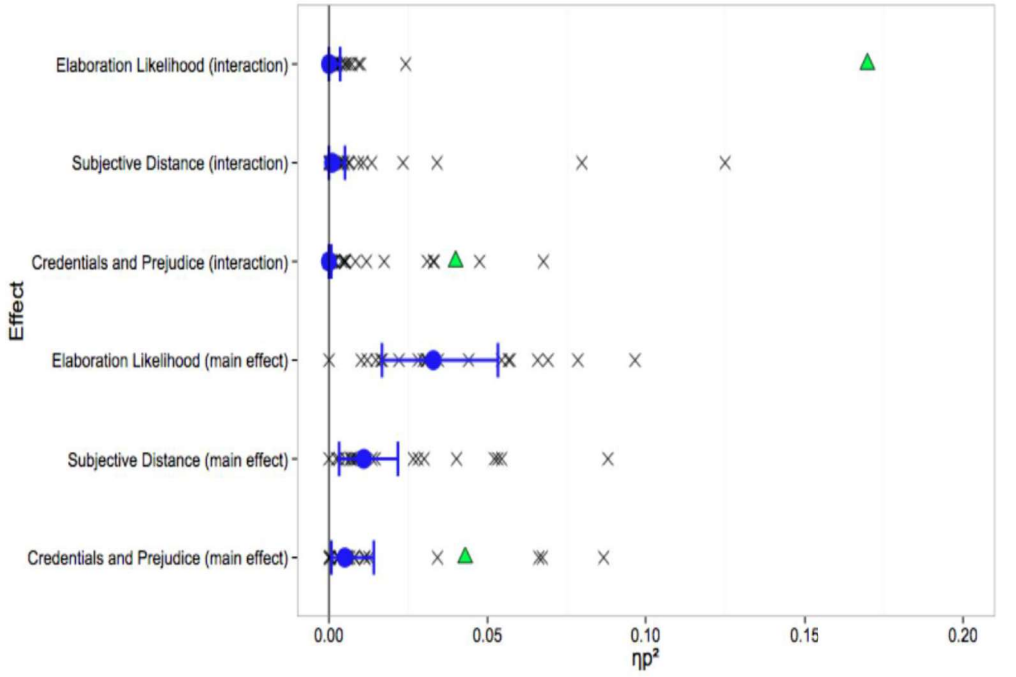
Many Labs (SP, 2014)



Many Labs 3 (2017)

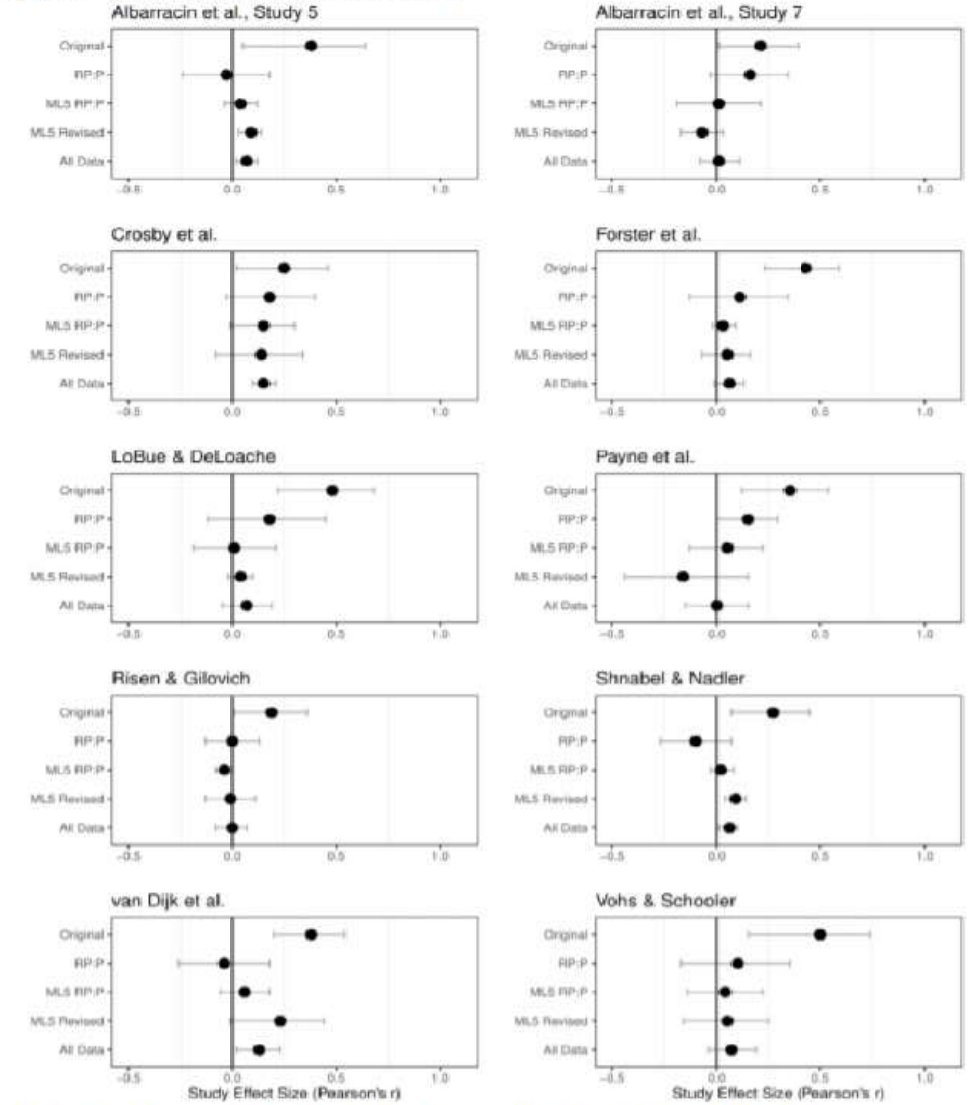
Figures 1a and 1b. Replication results organized by replication effect size, 1a for Cohen's *d* estimates, 1b for η_p^2 estimates. When available, the triangle indicates the effect size obtained in the original study (Stroop Effect and Elaboration Likelihood main effect estimate do not appear because they were very large, $d = 2.04$ and $\eta_p^2 = .59$ respectively). Large circles represent the aggregate effect size obtained across all participants. Error bars represent 99% noncentral confidence intervals around the effects. Small x's represent the effect sizes obtained within each site.





Many Labs 5 (2020)

Figure 2 - Effect sizes across study versions



Note: "All Data" represents a random effects meta-analytic estimate including the original study, the RP:P replication (OSC, 2015), and all Many Labs 5 data.

Many Labs 2 (2018)

Table 2. Summary of effect sizes, confidence intervals, and significance test counts across samples for each of the 28 studies

	Original Study		Repli		
				Global effects	
Effect	ES	95% CI	Median ES	ES	95% CI
Cohen's q Effect Size					
Disgust & Homophobia (Inbar et al., 2009)	0.70	.05, 1.36	0.03	0.05	.01, .10
Assimilation & Contrast (Schwarz et al., 1991)	0.48	.07, .88	-0.06	-0.07	-.12, -.02
Cohen's d Effect Size					
Correspondence Bias (Miyamoto & Kitayama, 2002) - WEIRD	2.47	1.46, 3.49	1.78	1.81	1.75, 1.88
Correspondence Bias (Miyamoto & Kitayama, 2002) - less WEIRD	0.74	-.12, 1.59	1.86	1.84	1.74, 1.94
Intentional Side Effects (Knobe, 2003)	1.45	.79, 2.77	1.94	1.75	1.70, 1.80
Trolley Dilemma 1 (Hauser et al., 2007)	2.50	2.22, 2.86	1.42	1.35	1.28, 1.41
False Consensus 1 (Ross et al., 1977)	0.99	0.24, 2.29	1.08	1.18	1.13, 1.23
Moral Typecasting (Gray & Wegner, 2009)	0.80	.31, 1.29	1.04	0.95	.91, 1.00
False Consensus 2 (Ross et al., 1977)	0.80	0.22, 1.87	0.89	0.95	.90, 1.00
Intuitive Reasoning (Norenzayan et al. 2002) - WEIRD	0.00	-0.15, .15	0.95	0.95	.90, 1.00
Intuitive Reasoning (Norenzayan et al. 2002) - less WEIRD	0.69	.24, 1.13	0.50	0.56	.46, .65
Less is Better (Hsee, 1998)	0.69	.24, 1.13	0.86	0.78	.74, .83
Direction & SES (Huang et al., 2014) - WEIRD	0.83	.37, 1.28	0.66	0.55	.49, .61
Direction & SES (Huang et al., 2014) - less WEIRD	-0.59	-.99, -.19	-0.10	0.03	-.05, .13
Framing (Tversky & Kahneman, 1981)	1.08	.71, 1.45	0.38	0.40	.35, .45
Moral Foundations (Graham et al., 2009)	0.52	.40, .63	0.23	0.29	.25, .34
Trolley Dilemma 2 (Hauser et al., 2007)	0.34	.26, .42	0.22	0.25	.20, .30
Tempting Fate (Risen & Gilovich, 2008)	0.39	.03, .75	0.23	0.18	.14, .22
Priming consumerism (Bauer et al., 2012)	0.87	.41, 1.34	0.16	0.12	.07, .17
Incidental Anchors (Critcher & Gilovich, 2008)	0.30	.02, .58	0.00	0.04	-.01, .09
Position & Power (Giessner & Schubert, 2007)	0.55	.05, 1.05	0.01	0.03	-.01, .08
Direction & Similarity (Tversky & Gati, 1978)	0.48	.16, .80	0.03	0.01	-.02, .04
Moral Cleansing (Zhong & Liljenquist, 2006)	1.02	.39, 2.44	0.00	0.00	-.05, .04
Structure & Goal-pursuit (Kay et al., 2014)	0.49	0.001, .973	-0.02	-0.02	-.07, .03
Social Value Orientation (Van Lange et al., 1997)	0.19	<.001, .47	0.06	-0.03	-.08, .02
Priming warmth affects climate beliefs (Zaval et al., 2014)	0.31	.03, .59	0.00	-0.03	-.09, .03
Incidental Disfluency (Alter et al., 2007)	0.63	-.004, 1.25	-0.07	-0.03	-.08, .01
SMS & Well-Being (Anderson et al., 2012)	0.57	.20, .93	-0.05	-0.04	-.09, -.004
Choosing or Rejecting (Shafir, 1993)	0.35	-.04, .68	-0.04	-0.13	-.18, -.09
Affect & Risk (Rottenstreich & Hsee, 2001)	0.74	<.001, 1.74	-0.06	-0.08	-.13, -.03
Actions are Choices (Savani et al. 2010) - WEIRD	0.08	-.33, .50	-0.24	-0.21	-.23, -.18
Actions are Choices (Savani et al. 2010) - less WEIRD	-0.65	-1.01, -.30	-0.14	-0.12	-.16, -.08

Registered Replication Report

Many Labs 2: Investigating Variation in Replicability Across Samples and Settings

Richard A. Klein[1], Michelangelo Vianello[2], Fred Hasselman[3,4], Byron G. Adams[5,6], Reginald B. Adams, Jr.[7], Sinan Alper[8]

Many Labs 2 (2018)

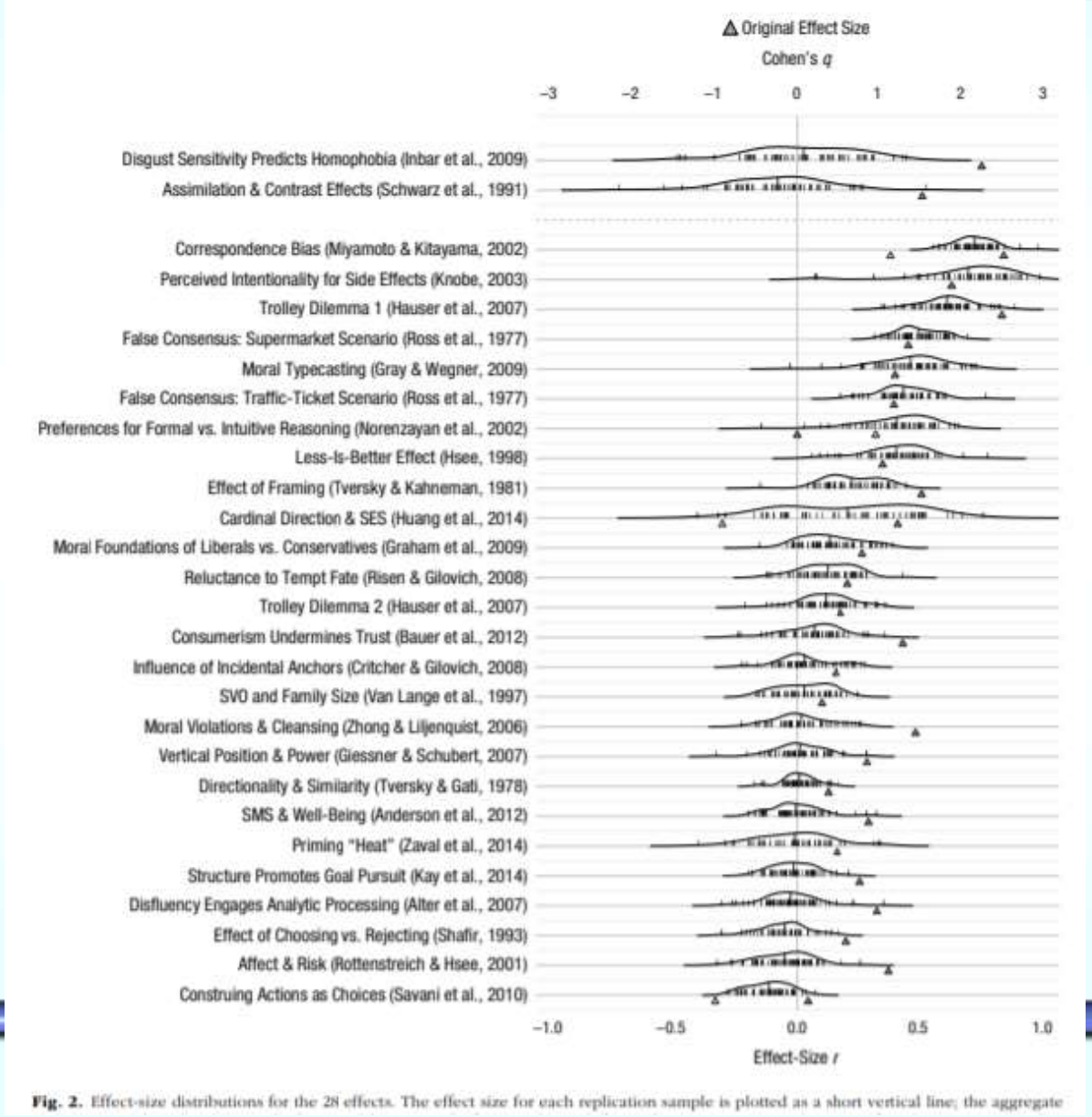


Fig. 2. Effect-size distributions for the 28 effects. The effect size for each replication sample is plotted as a short vertical line; the aggregate

14/28 (50%) sig. effect

1/28 (4%) weak effect

13/28 (46%) no or opposite effect

But sometimes most effects get replicated…

Personality (2019)

Psychological Science
2019, Vol. 30(5) 711–727
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797619831612
www.psychologicalscience.org/PS
SAGE

How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project

Christopher J. Soto
Department of Psychology, Colby College

66/76 (82%) sig. effects
ES: d=0.47 vs 0.61
Most original outcomes based on large samples rather than on searching for p<.05

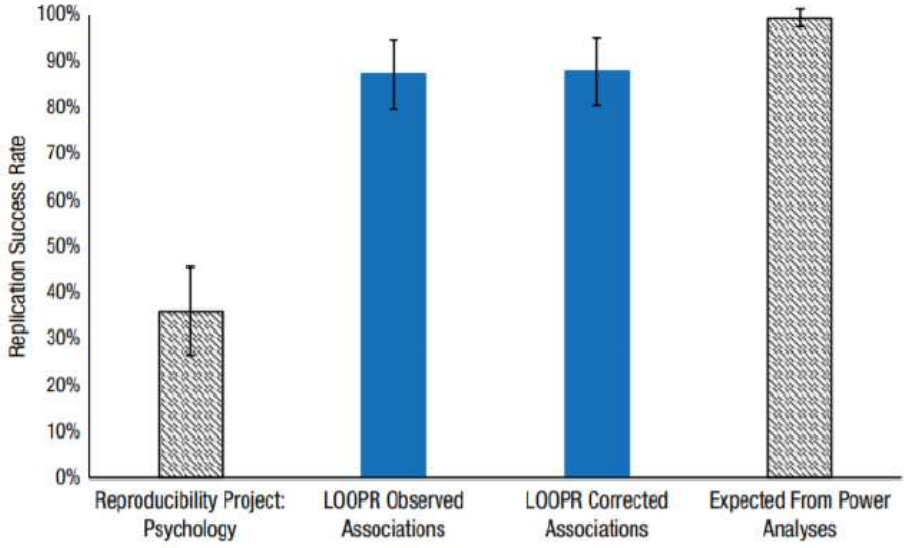


Fig. 1. Replication success rates obtained in the Life Outcomes of Personality Replication (LOOPR) Project, compared with the rate expected from power analyses of the original effect size and replication sample size and with the rate obtained in the Reproducibility Project: Psychology. A successful replication was defined as a statistically significant effect (i.e., two-tailed $p < .05$) in the hypothesized direction. Corrected associations were partially disattenuated to correct for the unreliability of abbreviated outcome measures. Error bars represent 95% confidence intervals.

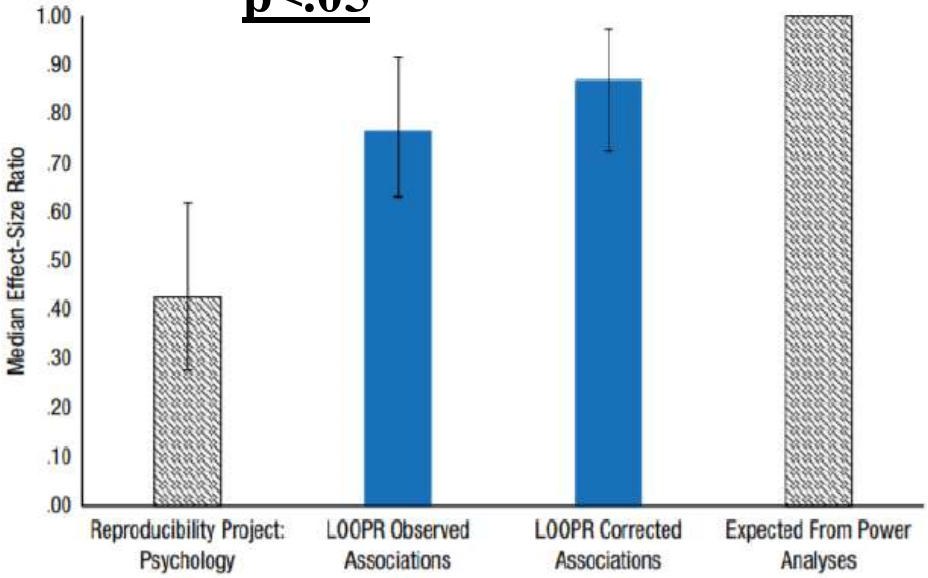


Fig. 2. Median effect-size ratios obtained in the Life Outcomes of Personality Replication (LOOPR) Project, compared with the ratio expected if all original effect sizes represented true effects and with the median ratio obtained in the Reproducibility Project: Psychology. Effect-size ratios were computed as the ratio of the *z*-transformed replication effect size to the transformed original effect size. Corrected associations were partially disattenuated to correct for the unreliability of abbreviated outcome measures. Error bars represent 95% confidence intervals.

Classic findings in Cognitive Psychology

Psychon Bull Rev (2018) 25:1968–1972
DOI 10.3758/s13423-017-1348-y

BRIEF REPORT

Participant Nonnaiveté and the reproducibility of cognitive psychology

Rolf A. Zwaan[1] · Diane Pecher[1] · Gabriele Paolacci[2] · Samantha Bouwmeester[1] · Peter Verkoeijen[1,3] · Katinka Dijkstra[1] · René Zeelenberg[1]

All nine effects replicated (highly significant effects)

<u>All within Ss studies</u>:

Table 1 Brief descriptions of and references to all replicated experiments

Number	Task	Description	Reference
1	*Simon task*	*Choice-reaction time task that measures spatial compatibility. Responses are faster when a visual target (a red square is presented on the left of the screen) is spatially compatible with the response (pressing the left button) than when the target is spatially incompatible with the response (presented on the right of the screen).*	Craft and Simon (1970)
2	*Flanker task*	*Response inhibition task in which relevant information is selected and inappropriate responses in a certain context are suppressed. Responses are faster for congruent trials in which compatible distractors flank a central target (AAAAA) than for incongruent trials in which incompatible distractors flank a central target (AAEAA).*	Eriksen and Eriksen (1974)
3	*Motor priming (a = masked, b = unmasked)*	*A task with a priming procedure in which responses to stimuli (arrow probes <<) are required that are primed by presented compatible (<<) or incompatible (>>) items. Responses are slower for compatible items when primes are masked but faster when primes are visible.*	Forster and Davis (1984)
4	*Spacing effect*	*Learning task in which learning (of words) is spaced over time. Recall of words is higher for spaced item repetitions with intervening items than for massed items immediately repeated after their first presentation.*	Greene (1989)
5	*False memories*	*Memory task that assesses false memory of recognition performance of items that have not been presented before in a word list but tend to be recognized as presented before because they are semantically related to the words in the list.*	Roediger and McDermott (1995)
6	*Serial position (a = primacy, b = recency)*	*Memory task that examines recall probability based on a word's position in a list. Recall is higher for the first and last words in the list and lowest for items in the middle of the list.*	Murdock (1962)
7	*Associative priming*	*Implicit memory task which requires a response to a target word that is preceded by prime word. Responses are faster when the prime is related than when the prime is unrelated.*	Meyer and Schvaneveldt (1971)
8	*Repetition priming (a = low frequency, b = high frequency)*	*Implicit memory task in which speed of response depends on previous exposure to an item and the word frequency of that item. Responses are faster for repeated than for new items. This repetition effect is larger for low frequency words than high frequency words.*	Forster and Davis (1984)
9	*Shape simulation*	*Sentence-verification task that requires a response on whether the object in a picture was present in the previous sentence. Yes responses are faster when the picture matches the implied shape mentioned in sentence than when it mismatches.*	Zwaan, Yaxley, and Stanfield (2002)

Some initial take-home messages

- Just because something has been shown once, it does not mean it is a consolidated fact
- There are robust and replicable effects and elusive effects
- Contextual variations (e.g., country, online vs. lab) seem to matter less than the effect as such
- Conceptual replication is not the same as direct replication, meta-analysis is not the same as multi-lab direct replication

Consequences

- Replicability is starting to have a stable place in Psychology (and in other sciences too)
- Beliefs in some effects is currently low (e.g., ego-depletion) but in others is high (e.g., anchoring)
- Research and publication standards are changing
- More awareness of false-positives and QRPs
- More attention to methodological issues
- More sophisticated statistical approaches

Why many effects are not replicated?

- A mix of different factors and possible explanations but two main factors
- Publication bias and **low power**
- Under these conditions, it is predictable that the literature will contain many false positives (results that seems significant but are not, see after) and artificially boosted effect sizes
- This is why they will be difficult to replicate
- We will come back to this issue towards the end

Power analysis

- Power analysis is an important tool in planning studies that test hypotheses
- We need first to understand in what context power analysis can be useful
- … and to refresh a few basic statistical concepts
- … and finally we can tackle power analysis