Replication-crisis and Publication-bias

Tadamasa Sawada

School of Psychology

01/Feb/2019

Wonder of Perception

Tadamasa Sawada
School of Psychology
01/Feb/2019

Question

Do I look different today?

Elliot et al. (2010)

Women find a man in a photograph more attractive, more sexually desirable, and having higher status when the man wears red clothes or when the photo's background is red.

Images from http://news.livedoor.com/article/image_detail/6513399/?img_id=3092585

Balcetis & Dunning (2010)

An object looks closer if you want it.

https://pixabay.com/en/structure-mud-dry-nature-864935/

Bem (2011)

A picture is hidden behind one of the gray squares above. Please guess which square is hiding the picture.

Bem (2011) found that the percent correct was significantly (p≈0.01) higher than chance-level (50%) if the pictures were **erotic**…

These results suggest that the humans have PSI: an ability to sense future events.

How strong is their evidence?

All of these studies were surprising, published in top journals, and sensationalized in many Science news sources.

The Null hypothesis was rejected in:

12 out of 12 experiments in Elliot et al. (2010, *JEP-General*)

5 out of   5 experiments in Balcetis & Dunning (2010, *Psy. Sci.*)

9 out of 10 experiments in Bem (2011, *J. Pers. Soc. Psy.*)

⋮

Note that the replication of an effect across multiple experiments (even within a single study) provided compelling evidence.

Reactions to Bem (2011)

Any problem in Methods? (Wagenmakers et al., 2011)

It is just speculative...

Any problem in Analysis? (Wagenmakers et al., 2011; Rouder & Morey, 2011)

Not really…

Isn’t it too good to be true? (Francis, 2012a)

Yes, his results are actually too good.

Too Good to Be True

Imagine you bought a special die that was made super-precise. It is so precise that you are guaranteed to see each face n times if you throw it $6n$ times. Is this too good to be true?

https://www.kickstarter.com/projects/124127689/precision-machined-dice

What Do You Want to Know from Statistics?

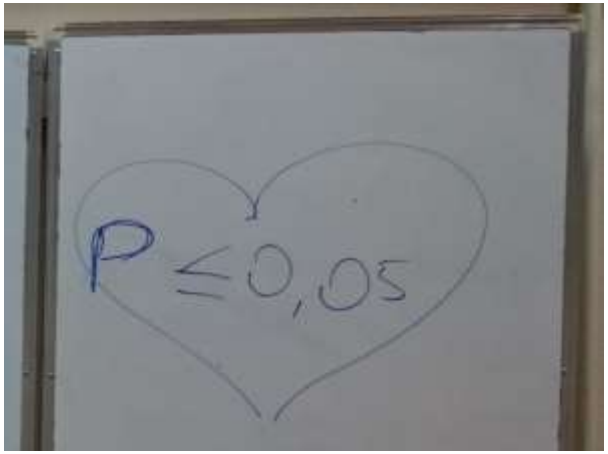
Imagine, you compare data measured under two conditions (G_1 and G_2). Statistical analysis suggests $G_1 > G_2$. Note that it is theoretically possible to observe such data even if the fact is $G_2 > G_1$ or $G_1 = G_2$. But, $G_1 > G_2$ is more ***likely*** than $G_2 > G_1$ and $G_1 = G_2$ from the data.

For empirical psychologists, statistics is the practical way to draw a border-line between two (or more than two) categories of data and to label them. Statisticians would hate this, but this is what the psychologists actually do.

Note well that a scientific theory is tentative (Ruse, 1981-1982). We can never deterministically know a fact from data but we can improve our understanding of the world from the data.

Categories of Data

Consider testing the effect of some treatment on cognitive performance. You compare the performance between two groups of participants: one with and one without the treatment. We analyze the results of the test by using a t-test and threshold the results: the effect of the treatment is significant if p < 0.05 and is not significant if p ≥ 0.05.



$p < 0.05$	The effect is significant
$p \geq 0.05$	The effect is not significant

Categories of Data

- **Type-1 error**: The effect does not exist but the statistical result suggests that it is (the null hypothesis H_0 is rejected).
- **Thpe-2 error**: The effect exists but the statistical result does not suggest that it does exist. This does “not” mean that H_0 is accepted.

	The effect Exists	The effect Does not exist
$p < 0.05$	Correct	Type-1 error (False positive)
$p \geq 0.05$	Type-2 error (False negative)	Correct

Categories of Data

- **Type-1 error**: The effect does not exist but the statistical result suggests that it is (the null hypothesis H_0 is rejected).
- **Thpe-2 error**: The effect exists but the statistical result does not suggest that it does exist. This does “not” mean that H_0 is accepted.
- **Power**: the Probability of observing a significant effect (rejecting H_0) with a given effect size for a given number of samples.

	The effect = $\boldsymbol{d}$ #Samples = N	The effect Exists	The effect Does not exist
$p < 0.05$	$\boldsymbol{Power}$	Correct	Type-1 error (False positive)
$p \geq 0.05$	$\boldsymbol{1 - Power}$	Type-2 error (False negative)	Correct

Statistical power and Frequency rejecting null-hypothesis

- **Power**: the probability that it correctly rejects the null hypothesis H_0 for a given effect size. The power also depends on the number of samples (subjects). The larger the sample number and the effect size are, the higher the power is.

If the effect is real but the power is low, then one would expect to fail to reject the null hypothesis with some frequency.

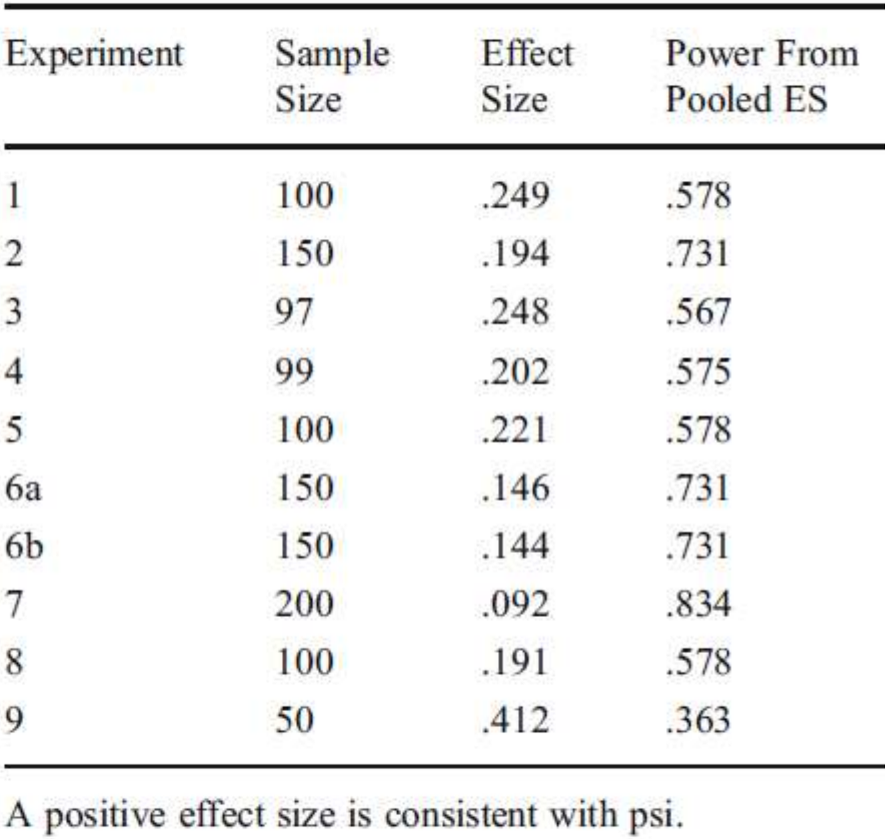
Then, how likely is it to observe such a compelling result (9 rejections of H_0 out of 10 experiments) as Bem (2011)?

Powers of the experiments in Bem (2011)

$$P(rejecting\ H_0\ in\ more\ than\ 9\ exp.\ out\ of\ 10)$$

$$=\quad p(rejecting\ H_0\ in\ all\ the\ 10\ exp.)$$

$$+\sum_{\boldsymbol{i}}^{10} p(rejecting\ H_0\ in\ every\ exp.\ except\ for\ the\ \boldsymbol{i}\text{–}th\ exp.)$$



Experiment	Sample Size	Effect Size	Power From Pooled ES
1	100	.249	.578
2	150	.194	.731
3	97	.248	.567
4	99	.202	.575
5	100	.221	.578
6a	150	.146	.731
6b	150	.144	.731
7	200	.092	.834
8	100	.191	.578
9	50	.412	.363

A positive effect size is consistent with psi.

$$P = 0.058 < 0.1$$

This 0.1 criterion came from:
Begg & Mazumdar (1994)
Ioannidis & Trikalinos (2007)
Stern, Gavaghan, & Egger (2000)

Too good to be true!

Publication bias

- Elliot et al. (2010): $P = 0.054 < 0.1$ (see Francis, 2013)
- Balcetis & Dunning (2010): $P = 0.076 < 0.1$ (see Francis, 2012b)
- Bem (2011): $P = 0.058 < 0.1$ (see Francis, 2012a)

⋮

These low probabilities suggest that they are too successful.

Publication bias!

Publication bias ≠ Effect does not exist

Note that a publication bias is not evidence that a phenomenon studied in multiple experiments is necessarily false.

However, if the bias is strong, it is hard to say whether the phenomenon is true or false.

Then, how different is this status from the status that existed before any experiment was run?

How common is Publication bias?

Publication bias was shown to exist:

in **82 % (36/44)** of empirical studies with ≥4 experiments published in ***Psychological Science*** between 2009 and 2012 (Francis, 2014)

in **83 % (15/18)** of empirical studies of Psychology with ≥4 experiments published in ***Science*** between 2005 and 2012 (Francis et al., 2014)

Psychological Science
RESEARCH, THEORY, & APPLICATION IN PSYCHOLOGY AND RELATED SCIENCES

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Science

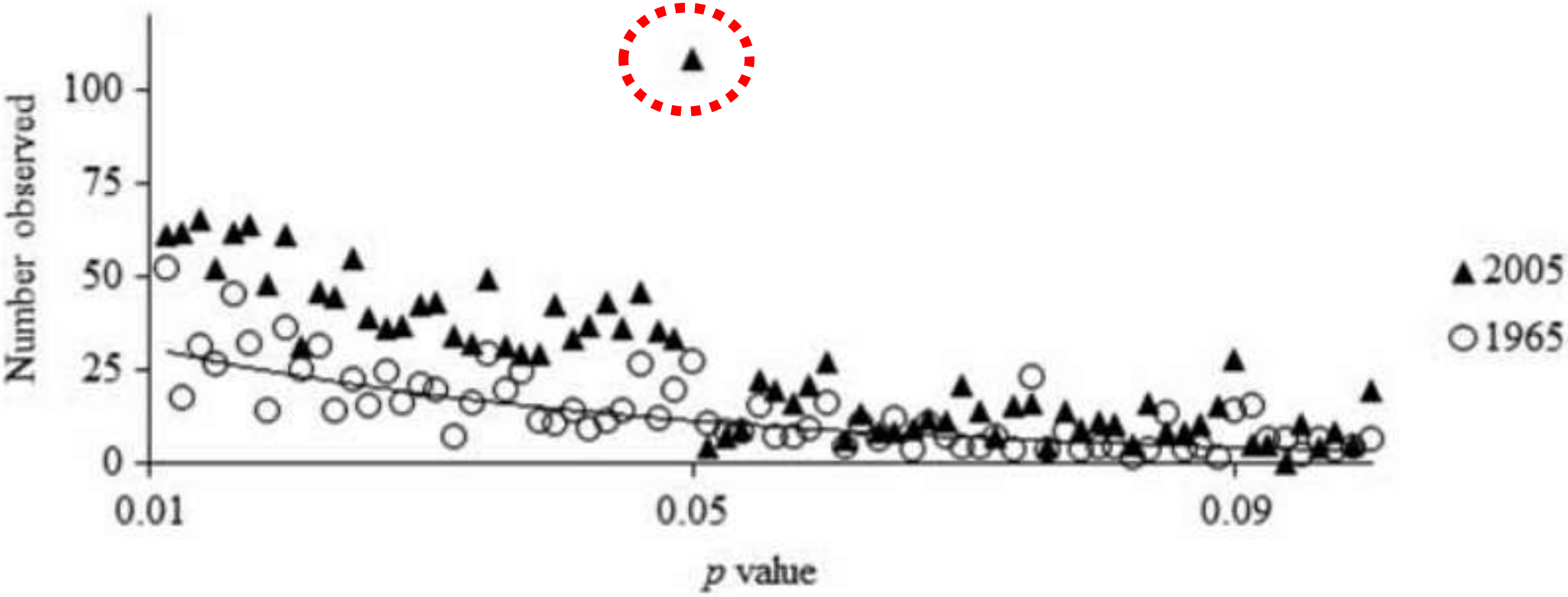
AAAS

Another evidence of Publication bias: *p*-value distribution

Distribution of ***p***-values in a Psychology journal

(Leggett, Thomas, Loetscher, & Nicholls, 2013)

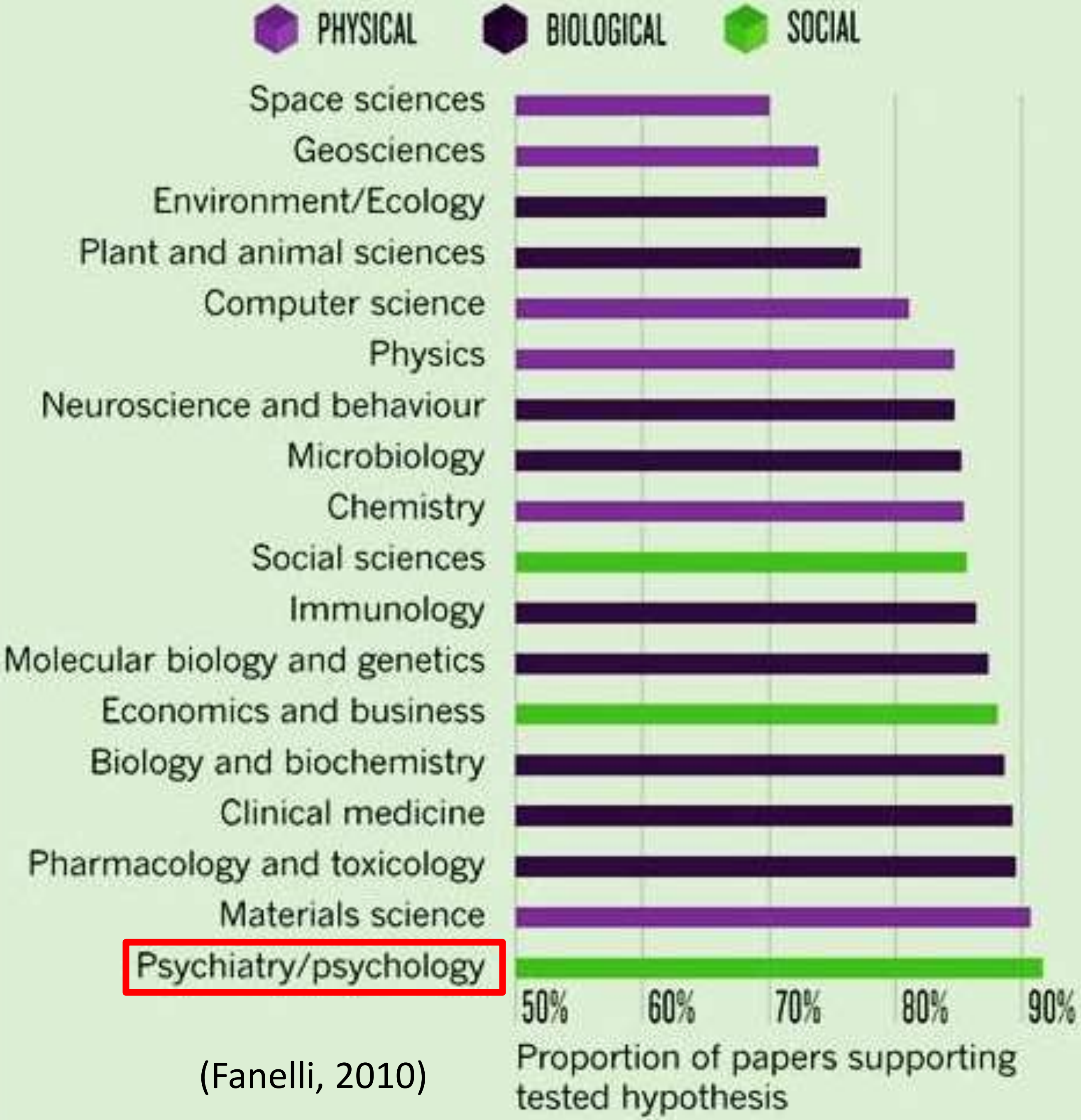
(See also Kühberger, Fritz, & Scherndl, 2014 for similar plots.)



The rule: $p < 0.05$

and the slogan:

Publish or Perish



(Fanelli, 2010)

Image from http://www.nature.com/news/replication-studies-bad-copy-1.10634

Replication Crisis

A group of 270 scientists (2015) who tried to replicate 97 experiments with positive results that were reported in 2008 issues of three prestigious journals (+ 3 experiments with non-significant results):

- *Psy. Sci.*
- *J. Pers. Soc. Psychol.*
- *J. Exp. Psy. Learn. Mem. Cogn.*

!!) 35/97 experiments could be replicated with $p < 0.05$.
!!) The average effect size among the replications (0.197) is less than half of that among the original studies (0.403).

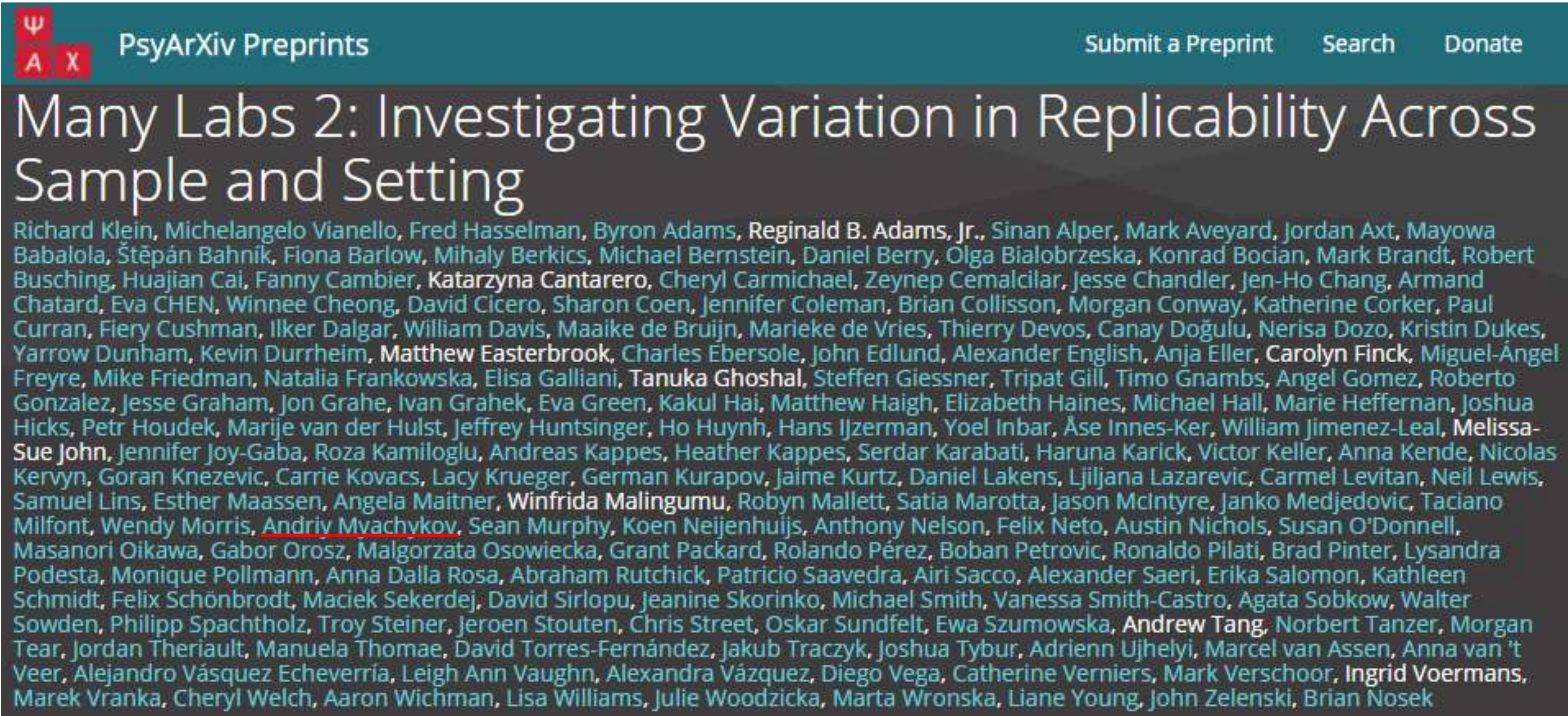
(Note that Gilbert et al. (2016) pointed out that many of the 97+3 experiments were not properly replicated; there were many methodological flaws. Actually, this is another problem of Psychology.)

Replication Crisis

A group of 187 scientists (2018) tried to replicate 28 experiments by using protocols that were peer-reviewed before any data were collected

!!) 15/28 experiments could be replicated with $p < 0.05$.

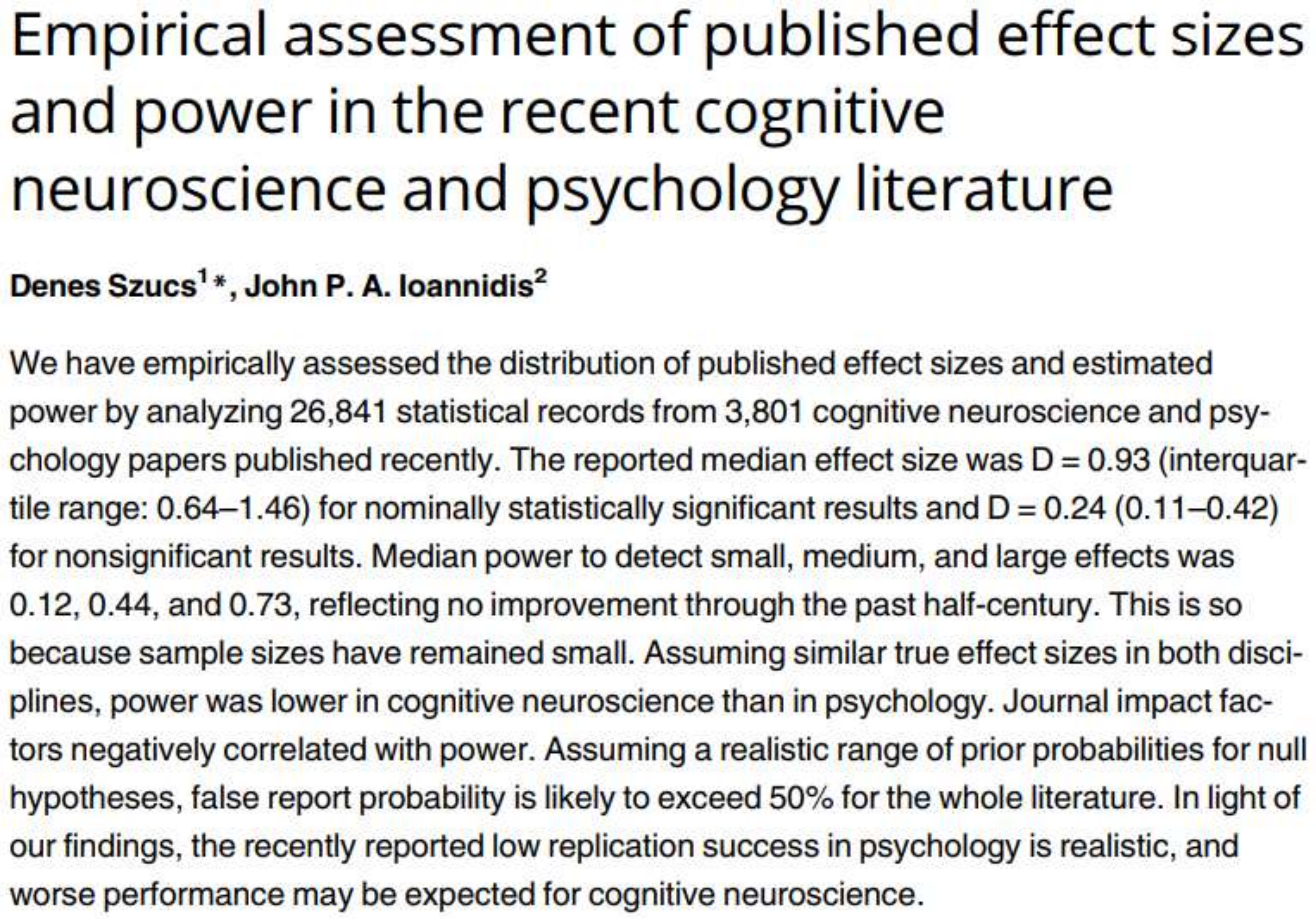
Could the criteria used for choosing the 28 experiments bias the results?



Many Labs 2: Investigating Variation in Replicability Across Sample and Setting

Richard Klein, Michelangelo Vianello, Fred Hasselman, Byron Adams, Reginald B. Adams, Jr., Sinan Alper, Mark Aveyard, Jordan Axt, Mayowa Babalola, Štěpán Bahník, Fiona Barlow, Mihaly Berkics, Michael Bernstein, Daniel Berry, Olga Bialobrzeska, Konrad Bocian, Mark Brandt, Robert Busching, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl Carmichael, Zeynep Cemalcilar, Jesse Chandler, Jen-Ho Chang, Armand Chatard, Eva CHEN, Winnee Cheong, David Cicero, Sharon Coen, Jennifer Coleman, Brian Collisson, Morgan Conway, Katherine Corker, Paul Curran, Fiery Cushman, Ilker Dalgar, William Davis, Maaike de Bruijn, Marieke de Vries, Thierry Devos, Canay Doğulu, Nerisa Dozo, Kristin Dukes, Yarrow Dunham, Kevin Durrheim, Matthew Easterbrook, Charles Ebersole, John Edlund, Alexander English, Anja Eller, Carolyn Finck, Miguel-Ángel Freyre, Mike Friedman, Natalia Frankowska, Elisa Galliani, Tanuka Ghoshal, Steffen Giessner, Tripat Gill, Timo Gnambs, Angel Gomez, Roberto Gonzalez, Jesse Graham, Jon Grahe, Ivan Grahek, Eva Green, Kakul Hai, Matthew Haigh, Elizabeth Haines, Michael Hall, Marie Heffernan, Joshua Hicks, Petr Houdek, Marije van der Hulst, Jeffrey Huntsinger, Ho Huynh, Hans IJzerman, Yoel Inbar, Åse Innes-Ker, William Jimenez-Leal, Melissa-Sue John, Jennifer Joy-Gaba, Roza Kamiloglu, Andreas Kappes, Heather Kappes, Serdar Karabati, Haruna Karick, Victor Keller, Anna Kende, Nicolas Kervyn, Goran Knezevic, Carrie Kovacs, Lacy Krueger, German Kurapov, Jaime Kurtz, Daniel Lakens, Ljiljana Lazarevic, Carmel Levitan, Neil Lewis, Samuel Lins, Esther Maassen, Angela Maitner, Winfrida Malingumu, Robyn Mallett, Satia Marotta, Jason McIntyre, Janko Medjedovic, Taciano Milfont, Wendy Morris, Andriy Myachykov, Sean Murphy, Koen Neijenhuijs, Anthony Nelson, Felix Neto, Austin Nichols, Susan O'Donnell, Masanori Oikawa, Gabor Orosz, Malgorzata Osowiecka, Grant Packard, Rolando Pérez, Boban Petrovic, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Monique Pollmann, Anna Dalla Rosa, Abraham Rutchick, Patricio Saavedra, Airi Sacco, Alexander Saeri, Erika Salomon, Kathleen Schmidt, Felix Schönbrodt, Maciek Sekerdej, David Sirlopu, Jeanine Skorinko, Michael Smith, Vanessa Smith-Castro, Agata Sobkow, Walter Sowden, Philipp Spachtholz, Troy Steiner, Jeroen Stouten, Chris Street, Oskar Sundfelt, Ewa Szumowska, Andrew Tang, Norbert Tanzer, Morgan Tear, Jordan Theriault, Manuela Thomae, David Torres-Fernández, Jakub Traczyk, Joshua Tybur, Adrienn Ujhelyi, Marcel van Assen, Anna van 't Veer, Alejandro Vásquez Echeverría, Leigh Ann Vaughn, Alexandra Vázquez, Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid Voermans, Marek Vranka, Cheryl Welch, Aaron Wichman, Lisa Williams, Julie Woodzicka, Marta Wronska, Liane Young, John Zelenski, Brian Nosek

Replication Crisis



Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature

Denes Szucs[1]*, John P. A. Ioannidis[2]

We have empirically assessed the distribution of published effect sizes and estimated power by analyzing 26,841 statistical records from 3,801 cognitive neuroscience and psychology papers published recently. The reported median effect size was D = 0.93 (interquartile range: 0.64–1.46) for nominally statistically significant results and D = 0.24 (0.11–0.42) for nonsignificant results. Median power to detect small, medium, and large effects was 0.12, 0.44, and 0.73, reflecting no improvement through the past half-century. This is so because sample sizes have remained small. Assuming similar true effect sizes in both disciplines, power was lower in cognitive neuroscience than in psychology. Journal impact factors negatively correlated with power. Assuming a realistic range of prior probabilities for null hypotheses, false report probability is likely to exceed 50% for the whole literature. In light of our findings, the recently reported low replication success in psychology is realistic, and worse performance may be expected for cognitive neuroscience.

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2000797

Behind the rule: File Drawer Problem

More than 50% of experimental data is unpublished (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986; Shadish, Doherty, & Montgomery, 1989). This is called the file drawer problem.

Image from https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

Behind the rule: Data-cooking and Playing with the Stats

Statisticians often receive requests for inappropriate analysis and reporting in Medicine (Wang et al., 2018). Note that 1/4 of them were asked for some kind of data-cooking.

Behind the rule: Arithmetic Errors

It is rather common that published articles have errors in their statistical results in Psychology (Veldkamp et al., 2014; Bakker & Wicherts, 2011). For example, 15% (39 out of 257) of articles have errors in statistical tests that can change their conclusions (Bakker & Wicherts, 2011).

$$1 + 1 = 3$$

Behind the rule: Decline Effect

Verbal-overshadowing is the suppression of visual memory that is caused by its verbal description (Schooler & Engstler-Schooler, 1990). This effect has been studied extensively in face recognition. This effect was initially easy to find but it has become increasingly difficult to replicate (Schooler, 2011). This is called the Decline effect.

There seems to be a publication bias ($p = 0.022 < 0.1$) among the studies of verbal-overshadowing (Francis, 2012a).

Image from http://www.whatsontv.co.uk/eastenders/episodes/sonias-grilled-about-pauline

Behind the rule: Decline Effect

- Regression toward the mean?

The *Sports Illustrated* cover jinx. A professional sports player/team who appeared on the front cover of a sports magazine tends to perform more poorly during the next season.

The player/team appeared on the cover because of their **outlying** performance during some period. In this case, we will subsequently observe a regression toward the mean. (Of course, this could happen also because other players/teams develop a counter plan against the star player/team.)

Behind the rule: Nonsense Projects

One third of published papers in Social Science are not cited within 5 years of their publication (Larivière & Gingras, 2008).

How many studies do report any real effect?

With the publication bias, the probability of a Type-1 error among published studies is not 5%. It may be inflated to a much larger number (Pashler & Harris, 2012).

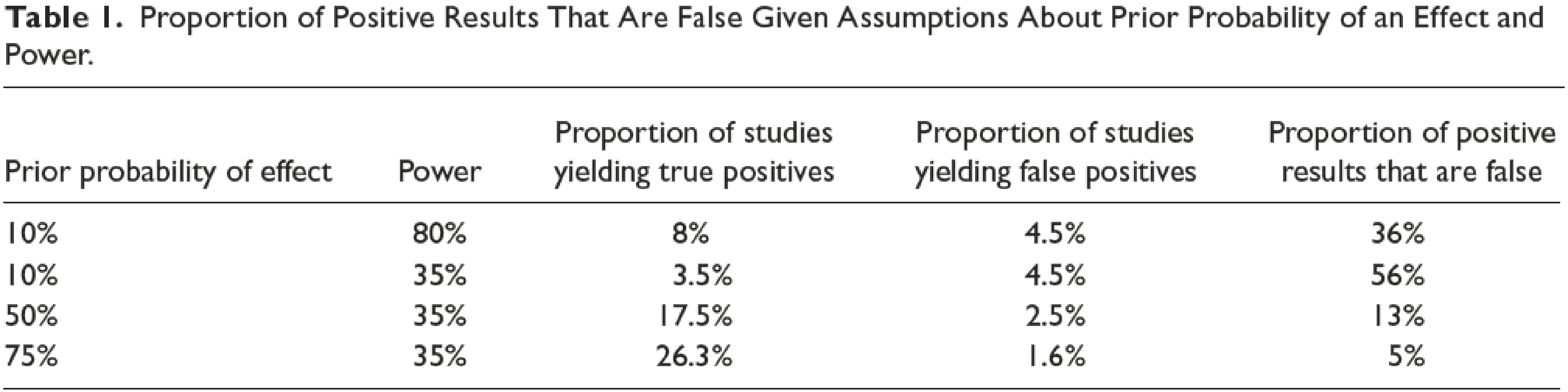


Table 1. Proportion of Positive Results That Are False Given Assumptions About Prior Probability of an Effect and Power.

Prior probability of effect	Power	Proportion of studies yielding true positives	Proportion of studies yielding false positives	Proportion of positive results that are false
10%	80%	8%	4.5%	36%
10%	35%	3.5%	4.5%	56%
50%	35%	17.5%	2.5%	13%
75%	35%	26.3%	1.6%	5%

a $\quad b$ $\quad c = a \times b$ $\quad d = (1-a) \times 0.05$ $\quad e = d/(c+d)$

What is a problem?

For example, the Mozart effect (Rauscher, Shaw, & Ky, 1993).

Despite increasingly definite null replications dating back to 1995 (e.g., Newman et al., 1995; Pietschnig, Voracek, & Formann, 2010), the Mozart effect persists in the popular imagination. Moreover, the Mozart effect was the basis of a statewide funding scheme in Georgia (Cromie, 1999), trademark applications (Campbell, 1997), and children's products; for instance, Amazon.co.uk lists hundreds of products that use the name 'The Mozart Effect', many touting the 'beneficial effects on the babies brain'. (from Bakker et al., 2013, Comments to Asendorpf et al. 2013)

What is a problem?

It can be true even for textbook-level ‘facts’.
For example, “behavioral priming” and “imitation of tongue gestures by young infants” are often introduced in Psychology textbooks. However, researchers have found that these studies are not well replicated and their current status is inconclusive (Bakker et al., 2013, comments to Asendorpf et al. 2013).

Even if some study is later rejected or becomes inconclusive, the original study remains “published”. Unless you are very careful, it is difficult to know that the study was subsequently rejected (see recent controversies about pseudosciene).

We like interesting studies but only if they are real.

What causes publication bias?

Publication practices of the society

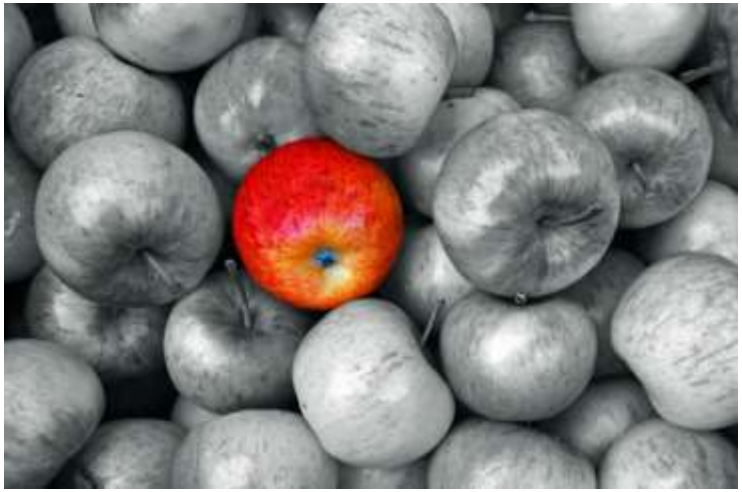
- Pressure on scientists to publish more papers
- Pressure on editors to accept more papers (on online journals)

What causes publication bias?

The poor research practices of individuals (p-hacking):

- Cherry picking
- Data peeking (Optional stopping)
- Multiple measurement
- HARKing (Hypothesizing After Results are Known, Kerr, 1998)
- Double dipping

Simmons et al. (2011) demonstrated that any false hypothesis (e.g. *listening to a children's song makes people feel older and listening to a song about older age makes people actually younger*) can be easily supported by empirical results by using these techniques.



What causes publication bias?

Psychological factor of individuals

- Confirmation bias
- Availability Heuristic (Tversky & Kahneman, 1973)
- Addiction to novelty, amazement, and surprise (Hume, 1748)
- Fallacy of autority

⋮

etc

Solutions?

- Data sharing (e.g. http://www.psychfiledrawer.org/)

Wicherts et al. (2006) requested data from authors of psychological studies published in APA journals and received data from only 64 out of 249 studies. Note that APA Ethical Principles include the principle of sharing data for re-analysis.

Solutions?

- Publishing non-significant results (e.g. PLOS ONE)
- Encouraging replications

Replication is not common in Psychology and, even when it is, it is not well-regarded and it can be biased (Bakker et al., 2012; Makel et al., 2012).

For example, Wiseman failed to replicate Bem's (2011) Psi effect. However, his study that had these negative results was rejected by the *Journal of Personality and Social Psychology* and *Psychological Science* because those journals do not publish replications. It was eventually published in the open-access journal *PLOS One* (2012). (http://www.apa.org/monitor/2013/02/results.aspx)

PLOS | ONE

Solutions?

- Revising the scientific process. For example:
  Step 1: Declaring protocols before conducting experiments
  Step 2: Reporting their results, regardless of out come
  (proposed by NIH National Institutes of Health Turning Discovery Into Health for medical studies)

Registered Replication Reports
(http://www.psychologicalscience.org/index.php/replication)

Archives of Scientific Psychology
(http://www.apa.org/pubs/journals/arc/index.aspx)

Open Science Framework
(https://osf.io/)

Solutions?

E.g. Cherry picking, Data peeking, Multiple measurement, and HARKing.

- Education

There is circumstantial evidence that some of the questionable research practices actually have been encouraged (Kerr, 1998; see also John, Loewenstein, & Prelec, 2012).

See also Lee (2018, BuzzFeedNews) about Wansink's case.

THE COMPLEAT
ACADEMIC
A Practical
Guide for the
Beginning Social Scientist
Edited by Mark P. Zanna & John M. Darley

Solutions?

- Education

Statistics, Probability, Methods, Ethics, Logics, and so on.

Solutions?

- Understanding properties of Psychology

  What scientists do in other field is not necessarily doable in Psychology. For example, data in Psychology tends to be much noisier than data in Physics.

How empirical results should look like?

(1) Flagship journals want “surprising” results.

“Extraordinary claims require extraordinary evidence”

By Carl Sagan

(2) Also, scientists are always under high pressure to publish papers in good journals. Computational simulation of evolution in Academia shows that the quality of studies in science deteriorates in such an environment (Smaldino & McElreath, 2016).

“The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”

By Donald T. Campbell (1976, taken from Smaldino & McElreath, 2016)

Solutions?

- Large effect size (high power)

Consider a study reporting a weak effect. This kind of study is hard to “***falsify***”.

Even if I try to replicate the study, it is hard to say whether my results support the effect or do not.
Case 1: If it is significant, my results support the original effect.
Case 2: If it is not significant, my results may still support the original effect because the power of the original weak effect is low and some non-significant results should be observed.

No matter whether the results are statistically significant or not, they can still support the original study. Remember, our resources are limited.

Solutions?

- Robust effect

… the replicator is closely copying the method set out in an earlier experiment, the original description must in some way be insufficient or otherwise defective.

Experimenters develop a sense, honed over many years, of how to use a method successfully. Much of this knowledge is implicit.

http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

Trivial details … could affect the results, and these subtleties never make it into methods sections.

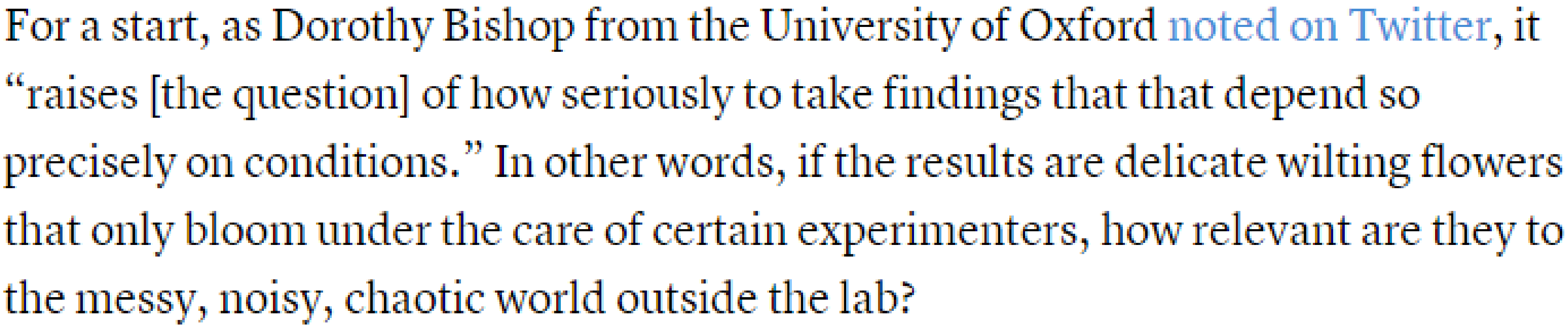
http://www.nature.com/news/replication-studies-bad-copy-1.10634

So, whenever you fail to replicate my study (visual perception), I have tons of excuses: moon phase, weather, latitude, #sunspots, gender of an experimenter, etc…

Solutions?

- Robust effect

If an effect is very delicate, does it ever affect us in any real environment? Do we need to pay attention to such delicate effects?



For a start, as Dorothy Bishop from the University of Oxford noted on Twitter, it “raises [the question] of how seriously to take findings that that depend so precisely on conditions.” In other words, if the results are delicate wilting flowers that only bloom under the care of certain experimenters, how relevant are they to the messy, noisy, chaotic world outside the lab?

www.theatlantic.com/notes/2015/09/sweeping-psychologys-problems-under-the-rug/403726/

Acknowledgement

Greg Francis

Zyg Pizlo