Misinformation and Disinformation in Statistical Methodology for Social Sciences: causes, consequences, and remedies

Giulio Giacomo Cantone, Venera Tomaselli

1 Introduction: the replicability of the Social Sciences

This paper concerns the prevalence and the causes of low replication rates in Social Sciences. The aim is to frame unintentional errors as scientific misinformation, and questionable research practices as disinformation. In Section 3 is presented Multiverse Analysis, which helps the assessment of the uncertainty about scientific claims and reduces false discoveries.

In order to introduce the topic of replication rate in Science, it is important to clarify the epistemological conditions to claim a scientific result to be replicated:

1. A scientific result consists of a claim A that is deduced through a procedure that can be reproduced by a third party (Goodman et al., 2016). A proper scientific result should be reported in an authoritative scientific venue, usually a peer-reviewed journal.
2. Others try to refute A by reproducing the same procedure on a different sample or adopting advanced but theoretically coherent alternative procedures on the original sample.
3. But these attempts fail: new results are not incompatible with A.

A replicated scientific theory is a collection of connected claims that are, for most, individually replicated (Lakatos, 1976; Schmidt, 2009). A replication rate is the rate of replicated results given a grouping variable: an author, an institution, or a scientific field. High replication rates are observed in exact sciences. Often, these replications are implicit: after a few successful experiments, a scientific theory is applied to more complex theories or technologies. The application of a theory is an implicit process of scientific replication (Feigenbaum and Levy, 1996). Methods of Social Sciences are not exact but probabilistic, harder to reproduce (e.g. due to changes in society), and applications into social policies are more nuanced than the vertical integration of natural sciences into technology.

Often claimed causal effects in Social Sciences are just statistical artifacts. Even meta-analyses are biased by so-called 'publication bias' (Nissen et al., 2016). It has been empirically demonstrated, indeed, that not significant estimates are less likely to be published in scientific venues (van Zwet and Cator, 2021). Prof. Breznau's research group provided the same dataset to 73 independent teams of quantitative social scientists, for a total of 161 people. He asked them to estimate the effect of immigration rates on public support for welfare-oriented political agenda. A sample of $n > 1,200$ estimate values for the effect has been drawn through this survey. Of the estimates, 25% were significantly negative, 17% significantly positive, and 57.7% of the times the specified model failed to reject the null hypothesis (Breznau et al., 2022). Impressively, based on this result, not only it is almost impossible to claim that a general effect exists, but even to fully deny it, because it is always possible to assert that an effect holds under specific conditions.

The U.S. Agency for Defense Advanced Research Projects (DARPA) understood the problem of traditional approaches for Meta-Analysis and Causal Inference and launched the Systematizing Confidence in Open Research and Evidence (SCORE) Project to understand how to

Giulio Giacomo Cantone, University of Catania, Italy, giulio.cantone@phd.unict.it, 0000-0001-7149-5213
Venera Tomaselli, University of Catania, Italy, venera.tomaselli@unict.it, 0000-0002-2287-7343

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giulio Giacomo Cantone, Venera Tomaselli, *Misinformation and Disinformation in Statistical Methodology for Social Sciences: causes, consequences, and remedies*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.10, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 53-58, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

predict if a study is deemed to fail to replicate. Preliminary findings have not been rosy: with exception of Economics, social scientists believe that their own fields produce more not replicable claims than replicable ones, i.e. there are more false discoveries than not. Economics seems to suffer of overconfidence in itself (Gordon et al., 2020). These results came after a large study led by Brian Nosek that attempted to replicate 100 claims in Psychology journals: less than half passed a replication attempt (OPEN SCIENCE COLLABORATION, 2015). Journals with high bibliometric scores do not perform better than other sources: evidence is in the direction of zero or negative correlation between bibliometric performances (e.g. journal impact factor) and replication rates (Szucs and Ioannidis, 2017; Brembs, 2018; Camerer et al., 2018).

2 Misinformation and disinformation

Ioannidis (2005) summarised predictors of low replication rates: small sample sizes, small effect sizes, and more than one hypothesis being tested on the same sample. On top of this, he stresses the incentives to look for novel findings instead of replication studies, too. He claims that papers on new theories are always more cited than their replication attempts, even when replication is not attained! This is a case of misinformation: inaccurate claims spread more than their corrections. Disinformation is a distinct phenomenon, where false claims are justified through a process of fabrication (West and Bergstrom, 2021). It is not necessary to report *fake data* to fabricate a fake result. The insidious alternative is to *omit* observed results. This behaviour is called "hacking the science" in the scientific community, by analogy with the method of *bruteforcing* many random combinations of inputs until a singular desired outcome is achieved by chance, e.g. hacking a password (Imbens, 2021).

2.1 Misinformation: is Duning-Kruger effect a statistical artifact?

It is commonly observed that the correlation between performance and self-assessment of performance is significantly negative. Since performance depends on skill, the theory of Duning-Kruger Effect (Kruger and Dunning, 1999) or DK, explains this correlation through the claim that unskilled people have a tendency to overestimate their own skills. The original study, with more than 8,000 citations, is foundational for modern Pedagogy. A concurrent to DK is the "better than average" theory (Krueger and Mueller, 2002), or BTA. It claims that all people have a tendency to self-assess their skills above the average, independently of their skill. These two theories can coexist but if BTA is true, then the DK effect is overestimated.

Consider the conservative case of two actors: one with a true skill score $x_1 = 40$ and the other with a true skill score $x_2 = 60$. Their average is $\bar{x} = 50$. Assume the claim of BTA: actor 1 and actor 2 have exactly the same model of assessment of self-score: they adopt the average plus an expected positive error ϵ^+. In this case, it holds

$$\left|x_1 - (\bar{x} + \epsilon^+)\right| > \left|x_2 - (\bar{x} + \epsilon^+)\right|, \forall \epsilon^+ \tag{1}$$

where $|x - (\bar{x} + \epsilon^+)|$ is the absolute error between true skill and self-assessed skill. It follows that: even with absolutely no cognitive differences between classes of actors (i.e. ϵ^+ is unique across actors), the less skilled actor has a larger absolute deviation. In this case, even if DK is not true, then the parameter ϵ^+ would induce a negative correlation. With few generalisations it is shown that any model that parameterises the self-assessed score to $\mu_X + \epsilon^+; \forall X : \{x_1, x_2, x_3, ..., x_n\}$ would lead into an artificial DK effect, even when DK is not true. The effect would hold even for normally distributed positive ϵ^+_{actor}.

A meta-analytical study that adopted advanced statistical techniques found that, given the observed scores in the literature, DK is likely to be a statistical artifice due to BTA (Gignac and

Zajenkowski, 2020). Another study reports only partial support for a true DK effect while confirming BTA (Jansen et al., 2021). Here no information has been concealed or fabricated. The authors did not adopt any questionable research practices. They lacked the correct specification of their null model.

2.2 Disinformation: six degrees of separation and even more

The expression "small world" refers to a network where a part of the connections happens with a uniform probability, and another part happens with a higher probability to form triadic closures (fully connected triangles of nodes). As emergent propriety, small world networks have a "characteristic average path length" L: for any given node in the network, any other node can be reached only by crossing paths with an expected length equal to L, independently by the number of nodes in the network.

Formation and structure of small-world networks have been described in the Watts-Strogatz model (Watts and Strogatz, 1998), but the description of this network goes back to Milgram (1967). Indeed, the implicit claim of Milgram is that in modern societies (pre-Internet) there is a characteristic path length L between human connections and that L is relatively short. Curiously, the paper with the experiment that originated the catchphrase "six degrees of separation" (Travers and Milgram, 1969) has been published only 2 years after a theoretical paper (Milgram, 1967) claiming the emergency of L in human societies. Together, the two papers collected more than 13.000 citations and, a rare case for a social science theory, they inspired new ideas not only in business (marketing, etc.) but also in engineering (transports, etc.).

It was a surprise for Judith Kleinfield (2002) to discover that the paper presenting the actual report of the *in vivo* experiment of the theory (Travers and Milgram, 1969) is actually poor in terms of statistical results. 296 participants have been recruited for the study. Their task was to send a document to one of their pre-existing social ties with the final aim that this document could reach a specific male broker in Boston. These 296 participants have been sampled across three populations: not brokers in Nebraska, brokers in Nebraska, and brokers in Boston.

This stratification would have been helpful if just enough documents reached their final destination: only 214 original participants sent the document and only 64 documents reached Boston's broker, after s stages. Among these 64, the observed average path length $l = 5.2$. The territorial variable was the only statistically significant. The number 6 (degrees of separation) is never explicitly mentioned, however, in footnote 4 the authors mention that they adjusted l through a not better specified marginal distribution of probabilities of reaching the final node at $s + 1$ stage (see paramter Q_i). In footnote 4, they claim a confidence interval for L between 5 and 7. Is there sufficient evidence for claiming that L exists? From the sample of not brokers from Nebraska, only 18 documents reached the destination, with $l = 5.7$. This result could be generalised to the U.S. population but the sample size would be small.

Kleinfield (2002) investigated Milgram's archives, looking for more. She only found concerning details:

- Milgram (1967) mentions a pilot study where a document has been received by a woman in only four days. Kleinfield found the pilot's report and concluded that Milgram picked an interesting anecdote but he never published more details about the pilot because it was a failure. Attrition in the pilot was so high to make meaningless the observed statistics. Q_i is never mentioned in the pilot.
- Travers and Milgram (1969) tried to alter the attrition rate in two ways: avoiding to recruit social outcasts and modifying the document from a single piece of paper to a "passport" in bright colours.
- She found an anonymous manuscript about a third attempt with inconsistent results.

2.3 *p*-hacking

The first case study falls under the category of 'misinformation within science' because it regards how the reputation of theories spreads within science even when a new model has been proven more consistent. The second case study is different: researchers concealed results from their own research because these were inconclusive toward their hypothesis. This is relatable to the case of so-called p-hacking of the level of significance α for rejection of the null hypothesis in statistical testing. p-hacking is a fraud because it omits to report the number of tests attempted before reaching a statistically significant result in data analysis (Simmons et al., 2011; Head et al., 2015). p-hacking is typically done in two ways:

1. Parallel p-hacking: many tests are arranged on different samples of the same population. Each sample has a minimal size but it is large enough to be deemed credible by the typical reader. Once a positive outcome is seen, no further test is necessary. In the reported result of the study, the number of tested samples is omitted and only the one associated with $p < \alpha$ is reported. As a reference: if the parameter of the effect size is equal to 0 and the null hypothesis of the test is true; with $\alpha = .05$, after 14 tests (Bernoulli trials of parameter α), the probability to see a $p < \alpha$ in at least a test is

$$\sum_{k=1}^{14} \alpha \cdot (1-\alpha)^{k-1} > .51 \tag{2}$$

   following the geometric distribution of the Bernoulli trials[1].
2. Sequential p-hacking: a multivariate dataset is collected and a hypothesis is formalised with a simple model. If the statistics of the model are not significant, then the specification of the model is trivially adjusted (e.g., control variables are added to the model, outliers are removed, data is pre-processed differently, etc.) until a random $p < \alpha$ is achieved. All of these operations are not reported. This is a fraudulent type of Hypothesising After Results are Known, or HARKing (Rubin, 2017).

3 Remedies: pre-registration and Multiverse Analysis

A possible remedy for *science hacking* is pre-registration, that is to record in a dedicated electronic archive an anonymous manuscript that details all the research questions and the methods of incoming research. This happens before the data collection, so in a peer-review authors can certify that their analysis is coherent with the original research design and that hypotheses are not drawn after knowing the sample statistics (Nosek et al., 2018). Pre-registration has two problems: (i) nothing prevents p-hacking a result, pre-registering its specification, then submitting the complete manuscript for peer-review (Yamada, 2018); (ii) it does not allow serendipitous discoveries incoherent with what is pre-registered (Simmons et al., 2021).

Looking back at the crowd-sourced estimation in Breznau et al. (2022), this approach is kindred to a meta-analytical paradigm called Multiverse Analysis: Gelman and Loken (2014) popularised the assumption that the robustness of a scientific model can be estimated through trivially altering its specification. They call "degrees of freedom of the researcher" the analytical choices in data analysis, e.g. the choice of a link function in binomial regression between *logit* and *probit*. Steegen et al. (2016) introduced the concept of the "multiverse" of a scientific claim. These degrees of freedom are the source of errors in estimation.

In particular, claims are formalised into models. Assuming that a *true parameter* θ of the model exists, given a dataset, exists a set $\Theta_j = \{\hat{\theta}_j\}$ of estimates from different j-specifications

[1]The equivalent command in R language is `pgeom(13,.05)`.

of the model such that each estimate $\hat{\theta}_j$ sufficiently close to θ and $\mathbf{E}(\hat{\theta}_j) = \theta$ holds. How to draw a sample that is representative of Θ_j in order to ascertain the uncertainty associated with the error of misspecification (model error)? Crowd-sourced estimation (Breznau et al., 2022) draws a random sample of specifications and estimates just by surveying experts. Instead, Multiverse Analysis draws a systemic (not random) sample $\hat{J}$ of specifications through mapping all the degrees of freedom of the researcher, e.g. inclusion/exclusion of control variables, operations in data pre-processing, modelling choices for overdispersion, etc. and combining them into $\hat{J}$, that is the multiversal sample of specifications or just the "multiverse".

Multiverse Analysis assumes that measures of variability in the observed multiversal estimates $\hat{\theta}_{j \in \hat{J}}$ are as much if not more informative than parametric or bootstrapped standard error or confidence intervals about the uncertainty involved in the estimation of θ (Young and Holsteen, 2017; Simonsohn et al., 2020). An interesting application of Multiverse Analysis is for checking the Janus effect (Patel et al., 2015), which is when in the same multiverse co-exist statistically significant $\hat{\theta}_j$, but with different signs. Janus Effect is a red flag in the sample of so-called parametric type S error (Gelman and Tuerlinckx, 2000).

References

Brembs, B. (2018). Prestigious Science Journals Struggle to Reach Even Average Reliability. *Frontiers in Human Neuroscience*, 12.

Breznau, N. et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119.

Camerer, C. F. et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Feigenbaum, S. and Levy, D. M. (1996). The technological obsolescence of scientific fraud. *Rationality and Society*, 8(3):261–276.

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6):460–466.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390.

Gignac, G. E. and Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, 80:101449.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12.

Gordon, M. et al. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, 7(7):200566.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.

Imbens, G. W. (2021). Statistical Significance, p-Values, and the Reporting of Uncertainty. *Journal of Economic Perspectives*, 35(3):157–174.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124.

Jansen, R. A., Rafferty, A. N., and Griffiths, T. L. (2021). A rational model of the Dunning-Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, 5(6):756–763.

Kleinfeld, J. S. (2002). The small world problem. *Society*, 39(2):61–66.

Krueger, J. and Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average

heuristic and statistical regression predict errors in estimates of own performance. *Journal of personality and social psychology*, 82(2):180.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

Lakatos, I. (1976). Falsification and the Methodology of Scientific Research Programmes. In Harding, S. G., editor, *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*, Synthese Library, pages 205–259. Springer Netherlands, Dordrecht.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Nissen, S. B., Magidson, T., Gross, K., and Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, 5:e21451. Publisher: eLife Sciences Publications, Ltd.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Patel, C. J., Burford, B., and Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058.

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, 21(4):308–320.

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, 13(2):90–100.

Simmons, J. P., Nelson, D., and Simonsohn, U. (2021). Pre-registration: Why and How. *Journal of Consumer Psychology*, 31(1):151–162.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214. Number: 11 Publisher: Nature Publishing Group.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci*, 11(5):702–712.

Szucs, D. and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3):e2000797.

Travers, J. and Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443.

van Zwet, E. W. and Cator, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4):437–452. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/stan.12241.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *nature*, 393(6684):440–442.

West, J. D. and Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117.

Yamada, Y. (2018). How to Crack Pre-registration: Toward Transparent and Open Science. *Frontiers in Psychology*, 9.

Young, C. and Holsteen, K. (2017). Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. *Sociological Methods & Research*, 46(1):3–40.