Meta-Psychology, 2024, vol 8, MP.2023.3987
https://doi.org/10.15626/MP.2023.3987
Article type: Commentary

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Rickard Cralsson
Reviewed by: Adrien Fillon, Lina Koppel
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/MX97J

The untrustworthy evidence in dishonesty research

František Bartoš
Department of Psychological Methods, University of Amsterdam

Replicable and reliable research is essential for cumulative science and its applications in practice. This article examines the quality of research on dishonesty using a sample of 286 hand-coded test statistics from 99 articles. Z-curve analysis indicates a low expected replication rate, a high proportion of missing studies, and an inflated false discovery risk. Test of insufficient variance (TIVA) finds that 11/61 articles with multiple test statistics contain results that are "too-good-to-be-true". Sensitivity analysis confirms the robustness of the findings. In conclusion, caution is advised when relying on or applying the existing literature on dishonesty.

Keywords: z-curve, TIVA, test statistics, statistical power, false positive risk

Introduction

The replicability of published literature has been, for a long time now, challenged by the replication crisis (Baker, 2016). Overestimated effect sizes, low statistical power, and inflated evidence were documented across a variety of disciplines (e.g., Bartoš, Maier, Wagenmakers, Nippold, et al., 2022; Bartoš et al., 2023; Fanelli, 2010; Fanelli et al., 2017; Ioannidis et al., 2017; Kvarven et al., 2020; Schwab et al., 2021; Stanley et al., 2018; van Aert et al., 2019). Research on dishonesty lies in the interdisciplinary area between social psychology and experimental economics, which exhibit varying replication rates (Camerer et al., 2016; Open Science Collaboration, 2015). While some recent replication attempts in dishonesty research have yielded positive results (e.g., Efendic et al., 2019; Prochazka et al., 2021; Wouda et al., 2017), other replication attempts failed to replicate previous findings (e.g., Kristal et al., 2020; van der Cruyssen et al., 2020; Verschuere et al., 2018). However, concerns regarding the trustworthiness of dishonesty research have recently escalated due to a series of data fraud allegations and article retractions (e.g., DataColada blog posts 98, 109, 110, and 110, http://datacolada.org; Proceedings of the National Academy of Sciences, 2021; Psychological Science, 2023a, 2023b).

Concerns about dishonesty research were already raised by Gerlach et al. (2019), who conducted so far the most comprehensive meta-analysis on dishonesty. Gerlach et al. (2019) identified 130 articles using at least one of four experimental paradigms (sender–receiver games, coin-flip tasks, die-roll tasks, and matrix tasks). Gerlach et al. (2019) used 'standardized report' measure (Abeler et al., 2019) to quantify the percentage of dishonest people in each setting and extracted data from 558 experiments covering 44,050 observations. Although the standardized report allowed Gerlach et al. (2019) to meaningfully meta-analyze results across different experimental settings, the transformed estimates and standard errors (or test statistics) provide less information about the publication process required for publication bias adjustment (e.g., Bartoš, Maier, Wagenmakers, Doucouliagos, & Stanley, 2022; Duval & Tweedie, 2000; Maier et al., 2023; Stanley & Doucouliagos, 2014; Vevea & Hedges, 1995). The loss of information regarding the publication process results from non-linear transformations applied to the originally observed estimates. In other words, since selection for statistical significance does not operate on the 'standardized report' itself, the 'standardized report' provides less information about the publication process. Despite this limitation, Gerlach et al. (2019) found a "substantial indication of publication bias in almost all measures of dishonest behavior" (p. 18), indicating that "the magnitude of dishonest behavior may be falsely estimated" (p. 18).

This study further examines the quality of studies included in Gerlach et al. (2019) by analyzing hand-coded focal test statistics using z-curve (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020) and test of insufficient variance (TIVA, Schimmack, 2014). The results suggest wide-spread selection for statistical significance, lacking statistical power, increased risk of false-positive results, and a significant proportion of too-good-to-be-true results.

Methods

See https://osf.io/kbqga/ for data and analysis scripts. The analysis was conducted in `R` (version 4.3, R Core Team, 2021) using the `zcurve` `R` package (version

2.3, Bartoš & Schimmack, 2020).

Data

I hand-coded test statistics of all focal hypothesis tests related to dishonest behavior (i.e., results that supported/opposed a hypothesized claim) from the 130 articles included in Gerlach et al. (2019).[1] Whenever possible, I used the originally reported test statistics, computed the test-statistics as the ratio of estimates and the corresponding standard errors, or used the reported *p*-values (original/recomputed test statistics are preferred as they suffer less from rounding errors). Out of the 130 articles, 99 articles contained 286 extractable test statistics (some articles reported only point estimates or stars). The vast majority of extracted test statistics were statistically significant; 193/286 test statistics were statistically significant on $\alpha = 0.05$, and 233/286 test statistics were statistically significant on $\alpha = 0.10$.

Z-curve

Z-curve is a statistical method for evaluating the quality of a heterogeneous set of studies. It approximates the distribution of statistically significant z-statistic in published studies by employing a mixture of truncated folded normal distributions. The mixture model provides a publication bias-corrected estimate of the mean statistical power of published studies (Brunner & Schimmack, 2020). The mean statistical power of published studies corresponds to the expected replication rate (ERR), the proportion of exact replication studies producing a statistically significant result in the same direction (but see Held et al., 2022; Ly et al., 2019; Pawel and Held, 2022 for other definitions and measures of replications).

Z-curve allows us to extrapolate beyond the sample of collected studies and provides an estimate of the mean power of all conducted, and possibly unreported, studies Bartoš and Schimmack (2022). The mean power of all conducted studies corresponds to the expected discovery rate (EDR), the proportion of conducted studies that were expected to be statistically significant. A discrepancy between the EDR and the observed discovery rate indicates selection for statistical significance (e.g., Rosenthal, 1979; Sterling, 1959)). Schimmack and Bartoš (n.d.) further demonstrated that EDR can be transformed into false discovery risk (FDR), the upper bound on false discovery rate—the proportion of false-positive results in the published literature (Sorić, 1989).

Test of Insufficient Variance

TIVA is a statistical method for identifying “too-good-to-be-true” results. It builds on the observation that p-values generated from studies with fixed power transform to z-statistics that follow an approximately normal distribution centered on a z-statistics corresponding to the studies’ power with variance equal to 1 (Schimmack, 2014).[2] Any heterogeneity in the power of the original studies leads to z-statistics following a mixture of the corresponding normal distributions. Consequently, the variance of such a mixture is necessarily larger than 1. TIVA uses this observation and tests whether the variance of observed z-statistics is lower than 1, indicating results that are unlikely to be obtained under unbiased sampling (Schimmack, 2014). While TIVA loses power to detect too-good-to-be-true results under heterogeneity, simulations studies showed it rarely exceeds the nominal error rate, making it a conservative test (Renkewitz & Keiner, 2019).

Sensitivity Analysis

One potential issue with hand-coding test statistics is a bias on the side of the coder (i.e., me). The coder may be more likely to code statistically significant test statistics than statistically non-significant ones. Such a bias would result in a negatively biased assessment of the literature. To address this potential issue, I performed a sensitivity analysis of the coding to assess the robustness of the results.

I assessed the robustness of each result by randomly replacing 5% to 100% (in steps of 5%) of the used test statistics. The randomly selected test statistics were replaced with test statistics simulated from well-powered and well-reported studies (power = 80%). Each random replacement was repeated 1000 times. In the limit (i.e., 100% replacement), the analyses should lead to unbiased results. However, if only a small proportion of the hand-coded test statistics required replacement to alter the conclusions significantly, it would indicate a lack of robustness in the presented findings.

Results

Z-curve

Figure 1 visualizes the z-curve conducted on all 286 extracted test statistics. The Figure highlights two findings (a) a prominent peak of statistically significant z-statistics just on the right side of a z-score corresponding to the statistical significance criterion ($z = 1.96$, vertical red line) and (b) a good fit of the z-curve model (blue

[1] In contrast to Gerlach et al. (2019), I coded test statistics also from experiments that did not use one of the four paradigms.

[2] A lower variance can be observed if the original statistical tests do not provide properly calibrated p-values or if nonconforming test statistics are coded as confirming.

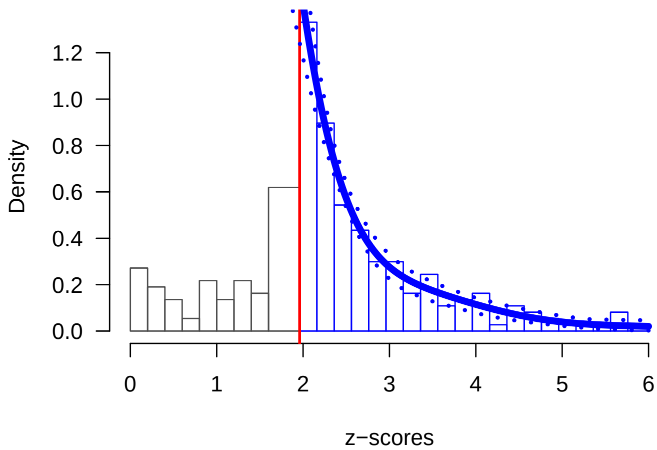

Figure 1

Z-curve highlights a substantial selection for statistical significance in studies on dishonesty.

line) to the observed distribution of statistically significant test statistics (blue histogram). The steep shape of the estimated z-curve is indicative of the very low expected replication rate, ERR = 0.378 [0.257, 0.491], while the "valley of missing statistically non-significant results" is reflected in the extremely low expected discovery rate, EDR = 0.082 [0.050, 0.191]. The EDR estimate is especially striking compared to the nine times larger observed discovery rate, ODR = 0.69 [0.63, 0.74]. The false discovery risk was very high, FDR = 0.0.590 [0.0.224, 1.000], but accompanied by a large degree of uncertainty due to the relatively small number of test statistics.

A secondary z-curve analysis was conducted to assess the sensitivity of the results to the potential non-independence of the test statistics (as 61 articles contributed more than one test statistic). The z-curve model was re-estimated while randomly selecting a single test statistic from each article (repeatedly to bootstrap CIs). The adjustment for non-independence did not meaningfully alter the results; ERR = 0.340 [0.238, 0.448], EDR = 0.075 [0.050, 0.144], FDR = 0.652 [0.312, 1.000].

Test of Insufficient Variance

All 61 articles with more than one test statistic were assessed by TIVA. The analysis revealed that 11/61 articles (18.0 [9.4, 30.0]%) reported results that were deemed "too-good-to-be-true" when testing the variance of the corresponding z-statistics against 1 with $\alpha = 0.05$.

Sensitivity Analysis

The sensitivity analyses showed that the presented results are robust to a considerably high percentage of potentially biased coding. Replacing even 25% of test statistics would not meaningfully alter the presented results. The detailed summary of sensitivity analysis to the potential coder bias is visualized in Figure 2. The x-axis depicts the proportion of replaced test statistics, and the y-axis depicts the target estimate (panel A: ERR, panel B: EDR, and panel C: FDR for z-curve, and panel D: percentage of too-good-to-be-true results for TIVA). The thick line corresponds to the median target estimate across the 1000 replications, and the thin lines correspond to a point-wise 95% quantile interval. The estimate and 95% CI at 0% of replaced test statistics correspond to the full sample estimate. Finally, the dotted red lines correspond to desirable estimates from well-powered and well-reported studies (i.e., ERR and EDR = 0.80, FDR of less than 5% although the FDR estimate converges to zero as all replacement studies are performed on true alternative hypotheses, and 5% of statistically significant TIVA results).

Discussion

The analysis of 286 test statistics from 99 articles included in Gerlach et al. (2019) revealed significant shortcomings in the quality of evidence in research on dishonesty. In particular, the examined set of studies suffered from low statistical power, high selection for statistical significance, inflated false positive rate, and many too-good-to-be-true results. Importantly, these conclusions do not speak about the quality of any particular article—there are undoubtedly many well-executed and well-reported studies. However, as a body of evidence, these articles fail to inspire trustworthiness.

The z-curve estimates for research on dishonesty are comparable to recently published estimates on motor learning benefits (McKay et al., 2023), effects of valenced odors on face perception and evaluation (Syrjänen et al., 2021), terror management theory (Chen et al., 2023), or system justification (Sotola & Credé, 2022). However, the general estimate for social psychology (Bartoš & Schimmack, 2022) or construal level theory (Maier et al., 2022) seems to be slightly better. Furthermore, top medical journals (Schimmack & Bartoš, n.d.), technology education research (Buckley et al., 2022), and organizational research (Gupta & Bosco, 2023), or tools and interventions for mitigating risks for gambling harm (McAuliffe et al., 2021), forced confabulation effect (Riesthuis et al., 2023), and social media use and self-esteem van Anen, 2022 seem to be of much higher quality. Finally, see Replicability-Index

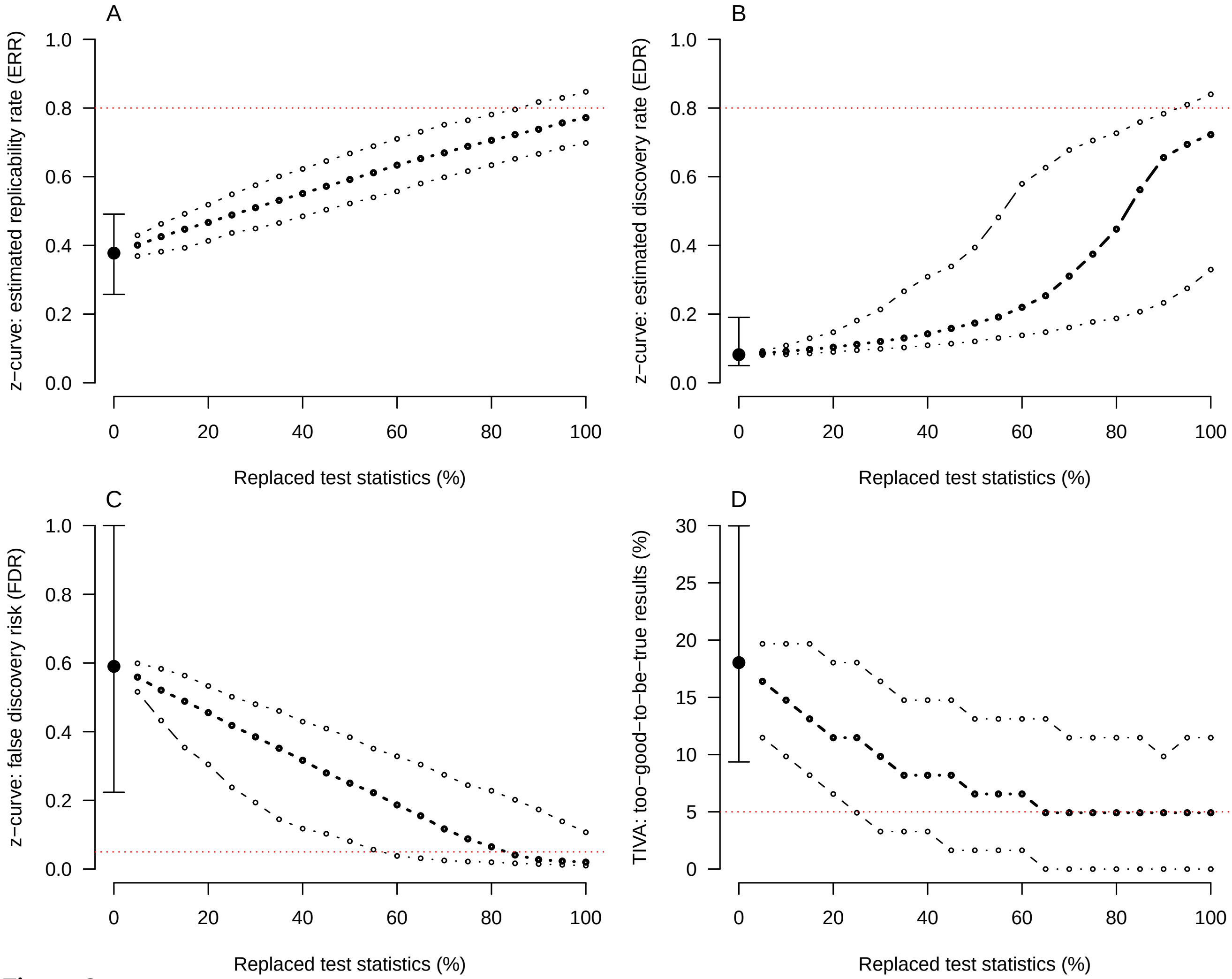

Figure 2

Sensitivity analysis to the hand-coding of the results shows robustness of results.

blog posts for psychological journals' specific z-curve estimates based on automatically extracted test statistics (e.g., https://replicationindex.com/2022/01/26/rr21/ for 2022 ratings).

Several limitations should be considered in the interpretation of these findings. While including all studies from Gerlach et al. (2019)'s meta-analysis removes the issue of "cherry-picking" articles, the generalizability of the presented conclusions might be limited. First, all examined articles employed one of the four most common experimental paradigms. Other designs or non-experimental studies might produce more reliable evidence. Second, all examined articles were published before 2019. There is reasonable hope that the ongoing methodological reforms improved the quality of the published literature. Third, all examined articles were coded by a single coder. However, the sensitivity analyses show that the results are robust to a large degree of biased coding.

In conclusion, consumers of the academic literature on dishonesty should be cautious when implementing or extending existing findings. While the trustworthiness of each study needs to be evaluated on an individual basis, there are some generic indicators of replicability. For example, studies with high sample sizes and large test statistics (e.g., $p < 0.001$) are more likely to replicate (Benjamin et al., 2018; Button et al., 2013; Fraley & Vazire, 2014). Furthermore, studies with open data can be re-analyzed with multiverse or many-analyst approaches which assess the robustness of the findings to the reported analytic choices (Gelman & Loken, 2013; Hoogeveen et al., 2023; Stern et al., 2019; Wagenmakers et al., 2022). Finally, new research should consider the registered reports format, which leads to highly credible evidence (Chambers, 2013; Chambers et al., 2015), practice modesty in interpreting results, and transparency in highlighting limitations (Hoekstra & Vazire, 2021).

Author Contact

František Bartoš; f.bartos96@gmail.com; Department of Psychological Methods, University of Amsterdam; ORCID: 0000-0002-0018-5573

Acknowledgment

The manuscript is based on work reported in the author's Bachelor's thesis 'Faktory asociované s podváděním' (2019).

The author would like to thank Maximilian Maier for his helpful comments on a previous version of the manuscript.

Data Availability Statement

See https://osf.io/kbqga/ for data and analysis scripts.

Conflict of Interest and Funding

No conflict of interest to report. This work was not supported by a specific grant.

Author Contributions

Mr. Bartoš solely contributed to all aspects of the research.

Open Science Practices

This article earned the Open Data and the Open Materials badge for preregistering the hypothesis and analysis before data collection, and for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115–1153. https://doi.org/10.3982/ECTA14673

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Bartoš, F. (2019). *Faktory asociované s podváděním* [Bachelor's thesis]. Univerzita Karlova, Filozofická fakulta. https://dspace.cuni.cz/handle/20.500.11956/107893

Bartoš, F., Maier, M., Shanks, D. R., Stanley, T., Sladekova, M., & Wagenmakers, E.-J. (2023). Meta-analyses in psychology often overestimate evidence for and size of effects. *Royal Society Open Science*, *10*(7), 1–12. https://doi.org/10.1098/rsos.230224

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. D. (2022). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, *14*(1), 99–116. https://doi.org/10.1002/jrsm.1594

Bartoš, F., Maier, M., Wagenmakers, E.-J., Nippold, F., Doucouliagos, H., Ioannidis, J. P. A., Otte, W. M., Sladekova, M., Deresssa, T. K., Bruns, S. B., Fanelli, D., & Stanley, T. D. (2022). *Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics*. https://doi.org/10.48550/arXiv.2208.12334

Bartoš, F., & Schimmack, U. (2020). zcurve: An R package for fitting z-curves [R package version 2.1.2]. https://CRAN.R-project.org/package=zcurve

Bartoš, F., & Schimmack, U. (2022). Z-curve. 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, *6*, 1–14. https://doi.org/10.15626/MP.2021.2720

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2018.874

Buckley, J., Hyland, T., & Seery, N. (2022). Estimating the replicability of technology education research. *International Journal of Technology and Design Education*, 1–22. https://doi.org/10.1007/s10798-022-09787-6

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

Chen, L., Benjamin, R., Guo, Y., Lai, A., & Heine, S. J. (2023). *Managing the terror of publication bias: A comprehensive p-curve analysis of the Terror Management Theory literature*. https://doi.org/10.21203/rs.3.rs-1254756/v1

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Efendic, E., Bartoš, F., Vranka, M. A., & Bahník, Š. (2019). *Unpacking the justifiability of dishonesty: Behavioral and process-tracing investigation* [Preprint at https://psyarxiv.com/rn85h].

Fanelli, D. (2010). "positive" results increase down the hierarchy of the sciences. *PloS One*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), 1–12. https://doi.org/10.1371/journal.pone.0109019

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gupta, A., & Bosco, F. (2023). Tempest in a teacup: An analysis of p-Hacking in organizational research. *PloS One*, *18*(2), e0281938. https://doi.org/10.1371/journal.pone.0281938

Held, L., Micheloud, C., & Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, *16*(2), 706–720. https://doi.org/10.1214/21-AOAS1502

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

Hoogeveen, S., Berkhout, S. W., Gronau, Q. F., Wagenmakers, E.-J., & Haaf, J. M. (2023). *Improving statistical analysis in team science: The case of a Bayesian multiverse of Many Labs 4*. https://doi.org/10.31234/osf.io/cb9er

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(605), F236–F265. https://doi.org/10.1111/ecoj.12461

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, *117*(13), 7103–7107. https://doi.org/10.1073/pnas.1911695117

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*(6), 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D., Harris, A. J., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31). https://doi.org/10.1073/pnas.2200300119

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2023). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods*, *28*(1), 107–122. 10 . 1037 / met0000405

McAuliffe, W. H., Edson, T. C., Louderback, E. R., LaRaja, A., & LaPlante, D. A. (2021). Responsible product design to mitigate excessive gambling: A scoping review and z-curve analysis of replicability. *PLoS One*, *16*(4), e0249926. https://doi.org/10.1371/journal.pone.0249926

McKay, B., Bacelar, M. F., Parma, J. O., Miller, M. W., & Carter, M. J. (2023). The combination of reporting bias and underpowered study designs has substantially exaggerated the motor learning benefits of self-controlled practice and enhanced expectancies: A meta-analysis. *International Review of Sport and Exercise Psychology*, 1–21. https://doi.org/10.1080/1750984X.2023.2207255

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science%20.aac4716

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 879–911. https://doi.org/10.1111/rssb.12491

Proceedings of the National Academy of Sciences. (2021). Retraction for 'Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end'. *Proceedings of the National Academy of Sciences*, *118*(38), 1–1. https://doi.org/10.1073/pnas.2115397118

Prochazka, J., Fedoseeva, Y., & Houdek, P. (2021). A field experiment on dishonesty: A registered replication of Azar et al.(2013). *Journal of Behavioral and Experimental Economics*, *90*, 101617. https://doi.org/10.1016/j.socec.2020.101617

Psychological Science. (2023a). Retraction notice to 'Evil genius? How dishonesty can lead to greater creativity'. *Psychological Science*, *34*(8), 947–947. https://doi.org/10.1177/09567976231187595

Psychological Science. (2023b). Retraction notice to 'The moral virtue of authenticity: How inauthenticity produces feelings of immorality and impurity.' *Psychological Science*, *34*(8), 948–948. https://doi.org/10.1177/09567976231187596

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Riesthuis, P., Otgaar, H., Bogaard, G., & Mangiulli, I. (2023). Factors affecting the forced confabulation effect: A meta-analysis of laboratory studies. *Memory*, *31*(5), 635–651. https://doi.org/10.1080/09658211.2023.2185931

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schimmack, U. (2014). The test of insufficient variance (TIVA): A new tool for the detection of questionable research practices [Blogpost at https://replicationindex.com/2014/12/30/tiva/]. https://replicationindex.com/2014/12/30/tiva/

Schimmack, U., & Bartoš, F. (n.d.). Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published p-values. *PLoS ONE*, *18*(8), 1–12. https://doi.org/10.1371/journal.pone.0290084

Schwab, S., Kreiliger, G., & Held, L. (2021). Assessing treatment effects and publication bias across different specialties in medicine: A meta-epidemiological study. *BMJ Open*, *11*(9), e045942. https://doi.org/10.1136/bmjopen-2020-045942

Sorić, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610. https://doi.org/10.2307/2289950

Sotola, L. K., & Credé, M. (2022). On the predicted replicability of two decades of experimental research on system justification: A z-curve analysis. *European Journal of Social Psychology*, *52*(5-6), 895–909. https://doi.org/10.1002/ejsp.2858

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi/10.1037/bul0000169

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad, Dinh, Grebe, Del Giudice, and Emery Thompson (2019). *Evolution and Human Behavior*, *40*(6), 517–525. https://doi.org/10.1016/j.evolhumbehav.2019.08.005

Syrjänen, E., Fischer, H., Liuzza, M. T., Lindholm, T., & Olofsson, J. K. (2021). A review of the effects of valenced odors on face perception and evaluation. *i-Perception*, *12*(2), 1–19. https://doi.org/10.1177/20416695211009552

van der Cruyssen, I., D'hondt, J., Meijer, E., & Verschuere, B. (2020). Does honesty require time? Two preregistered direct replications of experiment 2 of Shalvi, Eldar, and Bereby-Meyer (2012). *Psychological Science*, *31*(4), 460–467. https://doi.org/10.1177/0956797620903716

van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS One*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

van Anen, A. (2022). *How strong is our evidence? Evidential value and publication bias in research on social media use and self-esteem* [Master's thesis]. Tilburg University. http://arno.uvt.nl/show.cgi?fid=158963

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., et al. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317. https://doi.org/10.1177/2515245918781032

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wouda, J., Bijlstra, G., Frankenhuis, W. E., & Wigboldus, D. H. (2017). The collaborative roots of corruption? A replication of Weisel & Shalvi (2015). *Collabra: Psychology*, *3*, 1–3. https://doi.org/10.1525/collabra.97