ASSOCIATION FOR CONSUMER RESEARCH

Association for Consumer Research, University of Minnesota Duluth, 115 Chester Park, 31 West College Street Duluth, MN 55812

99% Impossible: a Valid, Or Falsifiable, Internal Meta-Analysis

Joachim Vosgerau, Bocconi University, Italy
Uri Simonsohn, University of Pennsylvania, USA
Leif D. Nelson, University of California Berkeley, USA
Joseph P. Simmons, University of Pennsylvania, USA

Researchers are increasingly relying on internal meta-analysis (IMA) to document cumulative evidence. We demonstrate that even minor violations of the assumptions underlying IMA invalidate its results—and once established—a false-positive IMA is practically impossible to falsify. We conclude that IMA should be used exclusively for pre-registered many-lab replications.

[to cite]:

Joachim Vosgerau, Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons (2019) ,"99% Impossible: a Valid, Or Falsifiable, Internal Meta-Analysis", in NA - Advances in Consumer Research Volume 47, eds. Rajesh Bagchi, Lauren Block, and Leonard Lee, Duluth, MN : Association for Consumer Research, Pages: 212-216.

[url]:

http://www.acrwebsite.org/volumes/2550879/volumes/v47/NA-47

Getting Wiser? New Insights into Consumer Research Practices

Chairs: Antonia Krefeld-Schwalb, University of Geneva, Switzerland
Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands
Benjamin Scheibehenne, University of Geneva, Switzerland

Paper #1: Using *P*-Curve to Assess Evidentiary Value from 10 Years of Published Literature

Leif Nelson, University of California, Berkeley, USA
Fausto Gonzalez, New York University, USA
Michael O'Donnell, Georgetown University, USA
Hannah Perfecto, Washington University in St. Louis, USA

Paper #2: 99% Impossible: A Valid, or Falsifiable, Internal Meta-Analysis.

Joachim Vosgerau, Bocconi University, Italy
Uri Simonsohn, ESADE, USA
Leif D. Nelson, University of California Berkeley, USA
Joseph Simmons, Wharton School, USA

Paper #3: The Price of Behavioral Research

Jason Roos, Erasmus University Rotterdam, The Netherlands
Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands

Paper #4: Tighter Nets for Smaller Fishes: Mapping the Development of Statistical Practices in Consumer Research Between 2011 and 2018

Antonia Krefeld-Schwalb, University of Geneva, Switzerland
Benjamin Scheibehenne, University of Geneva, Switzerland

SESSION OVERVIEW

In the last decade, behavioral researchers have faced increasing pressure to improve the methodology of data collection and analysis. A large survey documented the prevalence of research practices that are now considered "questionable" (e.g., stopping data collection when the desired results reached significance, John, Loewenstein, and Prelec, 2012). Similarly, several replication failures increased researchers' awareness of publication bias, file drawer effects, and other forms of selective reporting that can increase the prevalence of false positive findings in the literature (Ioannidis, 2005). A new wave of methodological research emerged, that revived old warnings (e.g., about too small sample sizes, Cohen, 1992) and proposed new standards for summarizing results (e.g., internal meta-analysis, McShane & Böckenholt, 2017) and new methods to assess the evidential value of a set of results (*p*-curve, Simonsohn, Nelson, and Simmons, 2014). Has consumer research become wiser as a result? How can it improve its methods?

The four papers in this session present new descriptive and normative insights into the state of the methodology of empirical consumer research. Nelson, Gonzalez, O'Donnell, and Perfecto will present a p-curve analysis of more than 400 articles in social psychology. This will yield insights into the robustness of phenomena that consumer researchers often build upon. Vosgerau, Simonsohn, Nelson, and Simmons argue that the validity of internal meta-analyses relies on assumptions (e.g., complete absence of selective reporting) that are often unrealistic. As a result, they show that relying on internal meta-analysis to establish the robustness of an effect can greatly inflate, rather than reduce, the rate of false positives. Roos and Paolacci leverage the transparency of online samples to investigate the extent to which monetary considerations are a substantial constraint to researchers' sampling decisions. They conducted an analysis of the MTurk study population before and after Amazon raised their commissions, and a coupon field experiment with researchers on Prolific. Results indicate that alleviating financial constraints would result in more studies and larger samples. Krefeld-Schwalb and Scheibehenne illustrate how selective reporting and statistical power in published consumer research has developed over time. By analyzing more than 900 articles in consumer research from 2011 to 2018, they found evidence that selective reporting has decreased. Together with an increase in sample size and decrease in effect size, these results suggest that the consumer research literature suffers less from false positive and false negative findings than it did in the past.

Altogether, these four papers contribute to our understanding of how individual consumer researchers and the field as-a-whole have navigated the methodological turmoil of the last few years and suggest ways to further strengthen our experimental research practices. We hope this session will trigger further interest in making consumer research wiser by improving its methodological practices.

Using *P*-Curve to Assess Evidentiary Value from 10 Years of Published Literature

EXTENDED ABSTRACT

P-curve is a tool that allows researchers to evaluate the evidentiary value in a given set of studies. The logic of *p*-curve is straightforward: because *p*-values are a conditional probability of observing a set of data (or data more extreme than what is observed) given that a null hypothesis is true (i.e. no effect), we know what the distribution of *p*-values should look like in the presence or absence of true effects.

If the null hypothesis is true, *p*-values will be uniformly distributed and show a flat line distribution with each value between 0 and 1 being equally likely. On the other hand, however, if an effect is true and the null hypothesis should be rejected, *p*-values should be strongly right-skewed with a spike at *p* approaching 0. The steepness of the curve is related to statistical power, with less power being associated with a flatter curve. This distribution of *p*-values will always occur when an effect is true, because if an effect truly does exist, the likelihood of finding large *p*-values is extremely small, and the bulk of *p*-values associated with statistical tests for a true effect will approach 0.

Finally, given the prevalence of *p*-hacking in published findings, *p*-curve is also a useful tool for identifying when *p*-hacking is likely present in a set of results. Because researchers are incentivized to report statistical tests with *p*-values at least below .05, if *p*-hacking is present in a set of studies, the distribution of *p*-values will be left-skewed with a spike approaching .05.

The simple logic of *p*-curve, the relationship between *p*-curve and statistical power, and the known distribution of *p*-values makes *p*-curve a powerful tool for researchers to assess how likely a set of studies (as in a paper, a journal, or a given research topic) are to contain evidentiary value for a true effect, and can be an important lodestar for researchers who are beginning to approach published findings at all levels.

To date, however, there does not yet exist a comprehensive database of *p*-values and there are few *p*-curves that cover a broad swathe of the literature. Indeed, while *p*-curving a literature is straightforward, it is an effortful process. Developing a *p*-curve disclosure table requires, at a minimum, reading a paper and reporting the research-

ers' hypothesis (with quotes from the paper), the study design, the key statistical result, statistical tests, and text from the paper verifying these elements. Although arduous, collecting this information is crucial for allowing authors of *p*-curved papers to easily assess the accuracy of the *p*-curve. In our paper, we are attempting to reduce these high startup costs associated with developing a large-scale *p*-curve database by systematically *p*-curving a decade's worth of findings in Section I of the *Journal of Personality and Social Psychology* (JPSP). We are also coding data beyond a typical *p*-curve disclosure table, such as keywords, whether the study excludes participants, specific experimental manipulations, whether the study tests mediation, and whether results are reported with and without covariates, in order to facilitate further metascientific analyses. We chose Section I of JPSP because it focuses on studies related to Attitudes and Social Cognition, contains the bulk of experiments in the journal, and is of the most direct relevance to Consumer Behavior researchers.

We have undertaken this endeavor as a means of developing a starting point for researchers who are interested in assessing the evidentiary value in a broad cross-section of psychological research that is relevant to scholars in marketing, management, social psychology, and many other behavioral fields. This *p*-curve database includes over 400 papers, each of which has been randomly assigned to be *p*-curved by two independent coders. The breadth of papers we have reviewed will allow researchers to assess entire streams of research within a given topic, evaluate trends in evidentiary value over time, assess whether specific manipulations or statistical results are associated with more or less evidentiary value, and so on. This will not only help get a sense of different literatures' reliability, but also help researchers developing new ideas decide which areas may be the most fruitful for building upon.

The development of this database is still ongoing, but the initial set of 10 years' worth of papers will be completed in Summer 2019. We plan to present a set of our most interesting, and consumer relevant findings from this database. We will also share the *p*-curve database as a public resource for consumer behavior researchers interested in evaluating the published research within *JPSP*.

99% Impossible: A Valid, or Falsifiable, Internal Meta-Analysis.

EXTENDED ABSTRACT

Internal meta-analysis involves statistically aggregating all studies reported in a paper, usually to examine whether the overall effect is statistically significant. Internal meta-analysis increases statistical power, potentially encouraging researchers to report more of their studies, particularly those that did not yield conventional levels of significance. These purported advantages – more statistical power and less "file-drawering" – have made internal meta-analysis popular.

We propose that internal meta-analyses are likely to have unintended and potentially catastrophic consequences for the credibility of published findings. The validity of internal meta-analysis hinges on the assumption that *none* of the analyzed findings were affected by selective reporting; the method is valid only if one analysis were conducted on each study, and only if every study was included in the internal meta-analysis.

We worry that some researchers may believe that the selective reporting of favorable analyses is immoral, a malevolent form of dishonesty. By this logic, saying that selective reporting is almost inevitable is like saying that almost all researchers are bad people. We believe that selective reporting of favorable analyses and studies is not immoral, but in most cases the inevitable consequence of (moral) human beings' tendency to interpret ambiguous information in ways that are consistent with their desires and beliefs (Kunda, 1990; Vazire, 2015).

To appreciate the near inevitability of this, consider what it would take to *not* do it. Researchers would either have to be indifferent to the outcome of their studies, or they would have to perfectly plan out in advance how many and which studies to run, which measures to analyze, how to score the measures, what sample sizes to use, which covariates to include, how exactly to deal with outliers or inattentive participants, etc. Motivated researchers who do not *perfectly* plan out their entire research project in advance will have to make *ex post* decisions about which studies to run, how to measure variables, which analyses use. And, because they are invested in the research project, they will make those decisions in ways that benefit them rather than in ways that harm them. Indeed, in the presence of desire and in the absence of perfect planning, some amount of selective reporting is virtually inevitable.

We show with simulations that small amounts of selective reporting can have dramatic consequences for internal meta-analyses. For example, the minimal selective reporting that inflates an *individual study's* false-positive rate to just 8% will inflate the false-positive rate of a 10-study internal meta-analysis to 82%!

Imagine that researchers only conduct internal meta-analyses on sets of studies that were individually pre-registered, so there is no *p*-hacking at all. Even under these exceptional circumstances, internal meta-analysis would be invalid if the decision about which studies to include in the meta-analysis was at all influenced by that study's results. If a researcher studying a false hypothesis needed 5 (out of 10) *individually* significant studies to successfully publish her result, she would have a one in 451,398 chance of succeeding. If the same researcher merely needed a significant *meta-analysis* of the best 5 out of 10 studies, she would have a 146,795 in 451,398 chance (or 32.5%) of succeeding. Ironically, internal meta-analysis *exacerbates* the consequences of the file-drawer problem, rather than alleviating it.

Finally, false-positive internal meta-analyses are prohibitively difficult to falsify, because if the original studies are distorted by selective reporting, the combination of original and replication studies will also be distorted by it. Using the previous example of a false-positive 10-study internal meta-analysis showing an effect, we proceeded to add 10 studies drawn under the null, re-running the meta-analyses now with 20 studies each (10 original and 10 replication studies). When all ten new replications had the same sample size as the original, 47% of false-positive internal meta-analyses remained significant. When all ten replications had 2.5 times the original sample size, still 30% of internal meta-analysis remained significant. Keep in mind that all of this assumes something extremely optimistic and unrealistic – that replicators could afford (or would bother) to replicate every single study in a meta-analysis and that others would judge all of those replication attempts to be of sufficient quality. Absent this wild assumption, we are left with the possibility that one cannot ever realistically attempt to falsify a false-positive internal meta-analysis.

Concluding, internal meta-analyses are valid only if (1) they exclusively contain studies that were properly pre-registered, (2) those pre-registrations were perfectly followed, and (3) the decision of whether to include a given study in an internal meta-analysis is made before *any* of those studies are run. These conditions are typically met in many-lab replication efforts, where the set of studies to be run is predetermined, and the exact design of each study is pre- predetermined also (see e.g., Alogna et al., 2014; Ebersole et al., 2016; Hag-

ger et al., 2016; Klein et al., 2014; McCarthy et al., 2018; O'Donnell et al., 2018; Verschuere et al., 2018; Wagenmakers et al., 2016).

Outside many-lab replication efforts, we recommend to never draw inferences about the existence of an effect from internal meta-analyses. We don't believe in the robustness of anchoring effects or motivated reasoning or preference projection because their findings have been meta-analyzed; we believe in them because the studies supporting them are well designed and because exact replications of these effects have been overwhelmingly successful. Scientific knowledge advances one replicable study at a time.

The Price of Behavioral Research

EXTENDED ABSTRACT

Experimental studies conducted on samples of human participants generate the bulk of evidence that drives behavioral consumer research forward. By collecting more evidence, whether through additional studies or larger samples, consumer researchers can measure behavior and test theories with greater precision. However, collecting evidence carries a monetary cost, and thus often entails a significant investment of scarce resources. How do researchers trade off monetary and scientific considerations? Two parallel phenomena make this long-standing question both pressing and answerable. First, the low replicability rate of findings in the behavioral sciences has revived old concerns about insufficient statistical power (e.g., Cohen 1992; Fraley and Vazire 2014; Maxwell 2004). It is important to understand whether sampling practices are bounded by financial constraints. Second, data collection in consumer research is moving away from physical laboratories in favor of online labor markets such as Amazon MTurk and Prolific (Goodman and Paolacci 2017). This removes many of the fixed costs of data collection, and thus make sampling decisions more dependent on the marginal cost of collecting additional observations. Moreover, the transparency of these marketplaces makes it possible to study researchers' behavior with greater precision.

Normatively, how should researchers' decisions about evidence to collect respond to the price of such evidence? The prescription of standard power analysis is straightforward. Sample sizes should be based on expected error rates and effect sizes, with no consideration for monetary costs. If the prescribed sample size happens to be compatible with the researcher's budget, then the study can be conducted. Otherwise, the study is not supposed to be carried out. In other words, a different price should normatively lead to conducting a different number of studies, but not to collecting differently sized samples. We investigated these possible patterns in two separate studies on two major suppliers of online participants.

In the first study, we analyzed whether researchers responded to Amazon's decision to raise the MTurk commissions in July 2015—factually increasing the price of data collection by 27%—by conducting fewer studies. Crawling data via the MTurk Tracker (Difallah et al. 2015), we analyzed surveys posted on MTurk in the two months before and after the price increase. We sought to measure the effect of the price increase on researchers who were already MTurk users. Hence, the analysis only considers activity among requesters who posted at least one survey or survey-based experiment on MTurk in the two months before the increase. The 2,292 requesters meeting these criteria posted 11,689 surveys during this four-month period.

We conducted a non-parametric change point analysis (Matteson and James 2014) on four quantities: the total number of surveys posted, the number of unique requesters associated with those surveys (some requesters post more than one survey per day), the average payment per survey (not the hourly rate), and the maximum time participants were given to complete their surveys. The algorithm identified two change points—June 27 (less than a week after the new prices were announced, $p = .014$) and July 21 (the day before the price increase, $p < .001$). These change points partition the data into three periods—*pre-announcement*, *post-announcement*, and *post-increase*.

To summarize results, prior to the announcement, an average of 82.5 (SE 5.1) requesters posted 135.3 (SE 20.7) surveys each day, offering an average reward of \$.60 (SE \$.02). After the announcement, the number of requesters and surveys did not meaningfully change, though the average reward increased by more than half to \$.96 (SE \$.06) per survey and grew more volatile. Most importantly, in the two months following the price increase, while average rewards fell to a value closer to the pre-announcement average, \$.72 (SE \$.02) per survey, the daily number of surveys and active requesters dropped by about half, to 43.4 (SE 2.4) requesters and 62.0 (SE 3.6) surveys. The maximum time allotted to surveys did not meaningfully vary over the four months. In sum, these results suggest that researchers reacted to the price increase by conducting fewer studies.

Importantly, the MTurk Tracker does not allow observing sample sizes. To understand whether sampling prices affect sample sizes, we conducted a preregistered field experiment on the entire population of researchers registered with Prolific who reside in the Netherlands ($N = 167$). In a between-participants design, researchers were either assigned or not to receiving a 15% discount off the total cost of their next study. Contrary to the prescription of standard power analysis, we found that researchers receiving the discount collected significantly larger samples ($M = 290$ vs. 116, medians = 240 vs. 70, Wilcoxon rank sum test $p = .0086$). Many compatible explanations exist for this effect: the discount may have affected sample sizes by influencing researchers' choice of which study to run next, prompting a significant alteration in the design or sample size of an already-planned study, or increasing the budget allocated to a study. Regardless of the mechanism, a short-term decrease in the cost of collecting evidence led researchers to accumulate far more evidence than their counterparts in the control.

The results from our second study suggest that sample sizes are affected by financial constraints, even in relatively cost-effective research environments. Critically, underpowered research has many causes, including underused or misinformed power analysis, and failing to recognize questionable research practices as such (e.g., Simmons, Nelson, and Simonsohn 2011). However, these results suggest that above and beyond these components, researchers might collect smaller sample sizes than they would if they were less financially constrained.

Taken together, the findings from these two studies illustrate how the production of evidence in the behavioral sciences does not depend exclusively on scientific considerations. On the contrary, researchers respond to changes in the opportunity costs of empirical studies by adjusting their sample sizes, and by conducting more or fewer studies. We will discuss the statistical, design, and institutional solutions available to researchers to trade off financial and scientific considerations more efficiently.

Tighter Nets for Smaller Fishes: Mapping the Development of Statistical Practices in Consumer Research Between 2011 and 2018

EXTENDED ABSTRACT

The replicability of empirical findings is a core aspect of (consumer) research (Popper, 1959). Recently, the need to improve the replicability of published findings has been called for across the

empirical social sciences, including consumer research (e.g. Ioannidis 2005). Towards this goal, several major journals have revised and strengthened their publishing policies, focusing more on the transparency and reproducibility of published research (e.g. Inman, Campbell, Kirmani, & Price, 2018; Pechmann, 2014). However, despite the awareness for the topic in the field, up to date, there has been no comprehensive evaluation of the replicability of published consumer research, nor how it developed or perhaps even improved in recent time. The present study provides a systematic review of the replicability of published consumer research. In particular, we investigated whether the replicability of consumer research has increased in recent years. Given the increasing discussion around the topic, such an improvement seems plausible.

The replicability of a statistical test depends on its statistical power which in turn depends on sample size and effect size. Once both are known, one can determine a distribution of *p*-values, i.e. the probability of observing different *p*-values. However, the effect size reported in published studies might not necessarily represent the true effect size of the statistical test, due to selective reporting and publishing of significant over non-significant findings (Duval & Tweedie, 2000; Egger, Smith, Schneider, & Minder, 2015; Ferguson & Brannick, 2012; Franco, Malhotra, & Simonovits, 2014; Rosenthal, 1979; Sterling, 1959). Together with other ways of selective reporting (Gelman & Loken, 2014), this increases the risk of overestimating the power of published studies and can increase the number of false positive findings in the literature and hence decrease the replicability of published findings.

One way to investigate the amount of selective reporting in published research is the analysis of the distribution of reported *p*-values (Simonsohn, Nelson, & Simmons, 2014). With any non-zero effect, the *p*-value distribution is right skewed, smaller *p*-values are more likely observed in a statistical test than higher p-values. If the effect size is zero, the distribution of *p*-values is uniform. In contrast, selective reporting, given that the effect size is zero, will produce a *p*-value distribution that is left skewed, below the level of significance. Thus, selective reporting distorts the distribution of p-values by increasing the proportion of values below the level of significance relative to the proportion that would be expected based on the observed statistical power.

Given the recent discussions in the social sciences on how to improve methodological and statistical practices in the field, we expected to observe an increase in the statistical power in the consumer research literature in the time period from 2011 to 2018. We further expected that publication bias decreased across that period. To test these hypotheses we focused on two commonly applied statistical tests in empirical consumer research, namely F-tests (i.e. analysis of variance) and t-tests. For both tests we analyzed the reported statistics across 971 articles published in the Journal of Consumer Research, the Journal of Consumer Psychology and the Journal of Marketing during that time period.

We first analyzed changes in the statistical post-hoc power over time. Given that sample sizes increased significantly from an average of 128 subjects in 2011 to on average 204 subjects in 2018, but effect sizes decreased across the same period from a median of .30 in 2011 to .22 in 2018, the average post-hoc power did not change. We further found no significant difference between the journals for any of the dependent variables.

Second, we tested whether the distribution of reported p-values changed in accordance with the distribution expected based on the reported power. On that account we analyzed the proportion of significant p-values in the articles as reported and expected based on the average reported power over time. We further separately investigated the development of p-values below and above a level of significance $\alpha = .005$ as proposed by Benjamin et al. (2017).

Despite that there was no decrease in statistical power over time, the reported proportion of significant findings, $p < .05$, decreased. However, the frequency of *p*-values in the intervals above and below the stricter level significance developed differently. *P*-values above the stricter level of significance, $.005 < p < .05$, decreased. Across all years and journals, *p*-values in this interval, $.005 < p < .05$, were yet more often reported than expected based on the power distribution. To the contrary, *p*-values below the stricter level of significance, $p < .005$, were less often reported than expected. Over time, however, the discrepancy of the expected and observed proportions of *p*-values has decreased.

One possible explanation for these results is that the increase in sample size over time indicate increased awareness for the need of higher powered studies. As a consequence, the decrease in effect size over time can be interpreted as a more exact estimation of the true sample sizes (Loken & Gelman, 2017; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Lane & Dunlap, 1978; Maxwell, 2004; Schmidt, 1992). Moreover, the reduction in discrepancy of the expected and the observed proportions indicates that selective reporting of significant findings in consumer research has decreased over time.

REFERENCES

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... & Buswell, K. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). Perspectives on Psychological Science, 9(5), 556-578.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E., ..., & Tingley, D. (2017). Redefine Statistical Significance. Human Nature Behavior, 1–18. http://doi.org/10.17605/OSF.IO/MKY9J

Cohen, J. (1992), A power primer. Psychological Bulletin, 112 (1), 155-159.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., Cudré-Mauroux, P. (2015) The dynamics of micro-task crowdsourcing: The case of Amazon MTurk in Proceedings of the 24th International Conference on World Wide Web, 238–247.

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. http://doi.org/10.1111/j.0006-341X.2000.00455.x

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67, 68-82.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (2015). Bias in meta-analysis detected by a simple, graphical test. Bmj, 14(7109), 1–16. http://doi.org/http://dx.doi.org/10.1136/bmj.315.7109.629

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychological Methods, 17(1), 120–128. http://doi.org/10.1037/a0024445

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, e109019

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences : Unlocking the file drawer. Science, 284(1983).

Gelman, A., & Loken, E. (2014). The statistical Crisis in science. American Scientist, 102(6), 460–465. http://doi.org/10.1511/2014.111.460

Goodman, J.K., Paolacci, G. (2017), Crowdsourcing consumer research, Journal of Consumer Research, 44 (1), 196-210.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. Perspectives on Psychological Science, 11, 546–573.

Inman, J. J., Campbell, M. C., Kirmani, A., & Price, L. L. (2018). Our Vision for the Journal of Consumer Research: It's All about the Consumer. Journal of Consumer Research, 44(5), 955–959. http://doi.org/10.1093/jcr/ucx123

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), 0696–0701. http://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychological Science, 23(5), 524–532. http://doi.org/10.1177/0956797611430953

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "many labs" replication project. Journal of Open Psychology Data, 2(1).

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480-498.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. Science, 355(6325), 584–585. http://doi.org/10.1126/science.aal3618

Matteson, D. S., James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association, 109, 334–345.

Maxwell, S. E. (2204). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological Methods, 9, 147 (2004).

Mcshane, B. B., & Böckenholt, U. (2017). Single Paper Meta-analysis: Benefits for Study Summary, Theory-testing, and Replicability. *Journal of Consumer Research*, *43*, ucw085. http://doi.org/10.1093/jcr/ucw085

Pechmann, C. C. (2014). Announcement Regarding the New Submission Guidelines at the Journal of Consumer Psychology. Retrieved from https://myscp.org/pdf/New Submission Requirements at JCP.pdf

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

Simmons, J. P., Nelson, L. D., Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. Journal of Experimental Psychology: General, 143(2), 534–547. http://doi.org/10.1037/a0033242

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. Retrieved from http://www.jstor.org/stable/2282137

Vazire, Simine (2015). This is what p-hacking looks like, blogpost, https://web.archive.org/web/20181018065906/http://sometimesimwrong.typepad.com/wrong/2015/02/this-is-what-p-hacking-looks-like.html, accessed February 19th 2018.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). Perspectives on Psychological Science, 11(6), 917-928.