Open Science 2019

March 12-14, 2019

Abstract Collection

Table of Contents

Authors:
Sophia Crüwell [1], Johnny van Doorn [1], Alexander Etz [2], Matthew C Makel [3], Hannah Moshontz [3], Jesse C Niebaum [4], Amy Orben [5], Sam Parsons [5], Michael Schulte-Mecklenbeck [6]

[1] University of Amsterdam; [2] University of California; [3] Duke University; [4] University of Colorado Boulder; [5] University of Oxford; [6] University of Bern

Title:
8 Easy Steps to Open Science: An Annotated Reading List

Session & Time:
"Open Science and its Impact on Scientific Practice" - March 12, 3:30pm - 4:00pm

Abstract:
Background
"Open Science" is an umbrella term used to refer to the concepts of openness, transparency, rigour, reproducibility, replicability, and accumulation of knowledge, which are considered fundamental features of science. In recent years, psychological scientists have begun to adopt reforms to make our work better align with these principles and address the current "credibility revolution" (Vazire, 2018). For example, the Society for the Improvement of Psychological Science (SIPS; https://improvingpsych.org/mission/) is a membership society dedicated specifically to improving the methods and practices of the field. The proposed open science reforms are largely a response to realisations that standard research practices undermine fundamental principles of good and open science (e.g., Ioannidis, 2005; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). Most scientists agree that there is a reproducibility crisis, at least to some extent (Baker, 2016). However, not all psychological scientists have adopted best practices designed to make science more reproducible (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; O'Boyle, Banks, & Gonzalez-Mulé, 2014). In part, this is because current incentive structures are misaligned with best practices (Bakker, van Dijk, & Wicherts, 2012; Higginson & Munafò, 2016). But there is also confusion and misinformation about what best practices are, whether and why they are necessary, and how to implement them (Houtkoop et al., 2018). In response, researchers have produced many excellent resources for each major facet of open science and methodological reforms that provide detailed instruction, context, and relevant empirical evidence. However, these articles are sometimes technical, are distributed across different journals and domains of psychology, and may be behind paywalls. Students and researchers with little background knowledge of open science may not easily find, access, or understand the resources that exist. Indeed, lack of information on the available resources and incentives for adopting best scientific practices have recently been identified as primary reasons for researchers in psychology not using best scientific practices (Washburn et al., 2018). Thus, an accessible and consolidated guide is needed to outline the best openly accessible resources related to best practices in (psychological) science.

Objectives and Approach
In this paper, we provide a comprehensive and concise introduction to open science practices and resources that can help students and researchers with no background

knowledge understand and implement best practices. Following the format of an annotated reading list introduced by Etz, Gronau, Dablander, Edelsbrunner, & Baribault (2018), we curate and summarise reviews, tutorials, and metascientific research related to eight topics: Open Science, Open Access, Open Data, Preregistration, Reproducible Analyses, Best Practices in Statistics, Replications, and Teaching Open Science. For each topic, we provide a summary of one publicly available, published, peer-reviewed article and suggest additional readings. Supporting a broader understanding of Open Science issues, this overview should enable researchers to engage with, improve, and implement current open, transparent, reproducible, replicable, and cumulative scientific practices.

Conclusions and Implications

One of the greatest barriers preventing psychological scientists from adopting open science practices is a lack of training and easily-accessible guidance. In this paper, we aim to address this barrier by providing a high-level, low-effort summary and overview of issues and solutions surrounding open science. Readers of all backgrounds can consult this text to understand the purpose of specific practices, obtain information about how to implement specific practices, and find more detailed resources.We hope that this paper will benefit individual researchers and the field as a whole. For all of the steps presented in this annotated reading list, the time taken to understand the issue and develop better practices is rewarded in orders of magnitude. On an individual level, time is ultimately saved, errors are reduced, and one's own research is improved through a greater adherence to openness and transparency. On a field-wide level, the more researchers invest effort in adopting these practices, the closer the field will come toward adhering to scientific norms and the values it claims to espouse.

Authors:
Edgar Erdfelder [1], Daniel W. Heck [1]

[1] University of Mannheim

Title:
Detecting Evidential Value and P-Hacking With the P-Curve Tool: A Word of Caution

Session & Time:
"Assessment of Publication Bias" - March 14, 11:00am - 11:30am

Abstract:
Simonsohn, Nelson, and Simmons (2014a) proposed *p*-curve – the distribution of statistically significant *p*-values for a set of studies – as a tool to assess the evidential value of these studies. They argued that, whereas right-skewed *p*-curves indicate true underlying effects, left-skewed *p*-curves indicate selective reporting of significant results from a much larger set of tests conducted on the same data when there is no true effect ("*p*-hacking").
We first review research that criticized the first claim by showing that null effects may indeed produce right-skewed *p*-curves under some conditions. We then question the second claim by showing that not only selective reporting but also selective non-reporting of significant results (e.g., of an ANCOVA for randomized 2-groups designs) due to a significant outcome of a more popular alternative test of the same hypothesis (e.g., a two-group *t*-test) may produce left-skewed *p*-curves, even if all studies included in a *p*-curve reflect true effects. Thus, although it is true that left-skewed *p*-curves indicate selection bias, it is possible that the bias is due to studies excluded from the *p*-curve rather than to those included in it. Hence, just as right-skewed *p*-curves do not necessarily imply evidential value, left-skewed *p*-curves do not necessarily imply *p*-hacking and absence of true effects in the studies involved.

Authors:
Malte Friese [1], Julius Frankenbach [1]

[1] Saarland University

Title:
P-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates

Session & Time:
"Consequences of Publication Bias on Statistical Evidence Generation" - March 13, 4:00pm - 4:30pm

Abstract:
Science depends on trustworthy evidence. The value of a biased scientific record is seriously compromised: Scientific progress is impeded and the public receives advice based on unreliable evidence with potentially far-reaching detrimental consequences. Meta-analysis enjoys high trust to reliably assess the state of the evidence in a given research field. However, meta-analytic effect size estimates may themselves be biased, threatening the validity and usefulness of meta-analyses to promote scientific progress. Here, we offer a large-scale simulation study to elucidate how *p*-hacking and publication bias distort meta-analytic effect size estimates under a broad array of circumstances reflecting the reality in varied research areas. The results reveal that, first, high levels of publication bias severely distort the cumulative evidence. Second, *p*-hacking and publication bias interact: at high and low levels of publication bias, *p*-hacking does little harm, but at medium levels of publication bias, *p*-hacking contributes considerable biasing effects. Third, while *p*-hacking can severely increase the rate of false positives, publication bias has the greater leverage to bias meta-analytic effect size estimates. A key implication is that policies in research institutions, funding agencies, and scientific journals need to make the prevention of publication bias a top priority to ensure a trustworthy evidence base.

Authors:
Tanja Fritz [1], Michael Kossmeier [1], Ulrich S Tran [1], Martin Voracek [1]

[1] University of Vienna

Title:
Reverberations of the Reproducibility Project Psychology: A content-based semantic citation analysis

Session & Time:
“Metascience” - March 12, 6:30pm - 7:00pm

Abstract:
The Reproducibility Project Psychology (RPP), published 2015 in the journal Science, is the largest and most-cited replication project in science to date. We employed content-based semantic citation analysis for an in-depth investigation of what the many citations the RPP received may mean. Based on its early reception (1000 citations from Google Scholar, 2015-17), resonance of the RPP was overwhelmingly positive (in 85% of citing papers), although the majority (74%) of RPP citations were superficial (not reporting any RPP statistics). Prevalence of open-science practices among RPP-citing papers was high; across science fields, the RPP was most frequently cited in psychological science (57% of all); and within this, in the subfields cognitive and social psychology and research methods. Methodology and metascience papers, along with empirical and replication studies, were the most frequent RPP-citing types of research. Altogether, the RPP marked a watershed in the 2010s debates on improving empirical research.

Authors:
Tobias Heycke [1], Lisa Spitzer [2]

[1] GESIS - Leibniz Institute for the Social Sciences; [2] University of Cologne

Title:
Screen recordings as a tool to document computer assisted data collection procedures

Session & Time:
"Open Science and its Impact on Scientific Practice" - March 12, 5:00pm - 5:30pm

Abstract:
In recent years, many scientific fields have re-discovered the need for replication of scientific studies. Famously, in a large replication project in psychology only approximately 40 % of the selected 100 studies were replicated by independent researchers (Open Science Collaboration, 2015). Additional (large-scale) replication attempts in psychology have confirmed initial findings that many published findings could not be replicated independently (e.g., Hagger et al., 2016; R. Klein et al., 2014).

Generally, one could propose two main reasons why a published finding cannot be replicated: First, the initial finding was merely a coincidence and the reported effect was not describing a natural phenomenon. Possible reasons for such a finding could be false positive statistical findings (which were possibly increased by p-hacking or other means of data massaging, Simmons, Nelson, & Simonsohn, 2011). Second, a finding might be based on a genuine phenomenon, but the effect was not replicated because essential details from the original experimental procedure were altered (assuming that the replication attempt had sufficient statistical power).
When replicating a study as close as possible, researchers "should strive to avoid making any kind of change or alteration" (Earp & Trafimow, 2015, p. 5). Direct replications might be highly informative, especially when the original results cannot be replicated in independent replications attempts. Specifically, not finding the original effect with a different method can easily be attributed to the method rather than the original effect (Doyen, Klein, Simons, & Cleeremans, 2014). Therefore, when a replication attempt does not succeed to find the original pattern of results, researchers might speculate whether the difference in the results could be due to subtle differences in the experimental procedure or critical changes that were introduced by the replicators (Freese & Peterson, 2017).

Practically, in (psychological) science post-hoc arguments can always be given to argue why studies could not replicate original findings. One might therefore be tempted to dismiss any post-hoc argument explaining why a replication attempt might have failed. However, what if effects indeed depend on small changes to the experimental procedure that are -so far- not understood by the scientific community? It should be considered highly problematic if small changes might lead to differences in the outcomes of an experiment, especially when researchers are potentially not aware of which details might be important and which not (see for example Alogna et al., 2014). If this is indeed the case, these details might therefore not

be reported in the written manuscript and we can only speculate if a non-replication might depend on one of these details. It is, however, simply not feasible to repeat an experiment with all combinations of potentially important methodological details.

One recommendation that appears to solve the above-mentioned problems, is to provide the research material to reviewers and post them publicly after publication (Asendorpf et al., 2013; Lindsay, 2017). The transparency and openness promotion (TOP) guidelines for example propose that “Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication” as the highest level of transparency of research materials (Nosek et al., 2015, p. 1424). In theory, uploading all materials to a public repository would solve most problems discussed previously.

However, in our opinion, there are a number of potential problems related to merely uploading experimental procedure scripts and material: First, one needs to possess the software the script was written for in order to run it. Unfortunately, many software solutions that are currently used are not freely available and the scripts can therefore not be run by every independent researcher. Second, even if one owns the software, or a freeware was used, the software version might have changed and the procedure might therefore look differently or the script does not run at all. Third, even if the software is still up to date and the researcher has access to it, he might not be acquainted with it and it may be time consuming to set up the script even when detailed instructions are provided. Even when running a replication with the original experimental script and material, it would be beneficial to know how the final procedure should look like. We therefore argue that there is still a need for better documentations of the research methods.

We propose that the experimental procedure should be recorded by means of screen capture and the video should be made available to others (e.g., by uploading it to a public repository). This way, the procedure is easy to access by reviewers, peers interested in the procedure and researchers interested in replicating the work. Importantly, screen recordings will not be affected by software changes, that produce a different look with the same experimental script. Additionally, researchers do not need to acquire and set up software solutions in order to have a look at the procedure that is likely more detailed than the description in the (published) manuscript. Especially referees in the peer review process would benefit highly from this documentation of the research method to inform themselves about the experimental procedure.

We have therefore created a tutorial on the open source screen recording software OBS (osf.io/3twe9). We would like to engage researchers in a discussion on possibilities to better document their experimental procedures and see this tutorial as a first starting point.

Authors:
Johannes Hönekopp [1], Audrey Linden [1]

[1] Northumbria University

Title:
How do questionable research practices affect inferences of heterogeneity? A computer simulation.

Session & Time:
"Consequences of Publication Bias on Statistical Evidence Generation" - March 13, 4:30pm - 5:00pm

Abstract:
Background
Heterogeneity reflects to what extent the results of a body of studies investigating the same question disagree over and above the effects of sampling error. Large heterogeneity means that effect sizes differ much more across studies than expected by the vagaries of sampling alone. Previously, we investigated heterogeneity in meta-analyses (which we might think of as consisting of a number of conceptual replications) and in multiple close replications (e.g. Many Labs). We found that heterogeneity tends to be very large in conceptual replications but quite small in close replications.
However, it is unclear to what extent publication bias and questionable research practices (QRPs) might distort the heterogeneity that we observed in meta-analyses. It is well known that publication bias and QRPs inflate effect sizes in the published literature. There might be some effect on heterogeneity as well.

Objectives
To understand to what extent we can trust the levels of heterogeneity empirically observed in meta-analyses.

Research question(s) and/or hypothesis/es
How do publication bias and wide-spread QRPs affect the observed heterogeneity in meta-analyses?

Method/Approach
We ran computer simulations in R. All considered between-subjects experiments in which means for experimental and control groups were contrasted. Multiple studies were run, published (or not) and (if published) summarised in a meta-analysis. We investigated the following factors in a fully crossed design: true effect size; true heterogeneity; number of studies per meta-analysis; strength of publication bias; QRP environment. The latter was defined by the extent to which (simulated) researchers engaged in the following: optional reporting from multiple dependent variables; optional stopping in participant recruitment; optional use of a moderator variable; and optional outlier removal.

Results/Findings

Overall, publication bias and QRPs tend to moderately inflate observed heterogeneity. However, this bias appears small compared to levels of observed heterogeneity.

Conclusions and implications (expected)
High levels of heterogeneity observed in meta-analyses need to be taken seriously, and cannot be conveniently explained as artefacts arising from publication bias and QRPs. We discuss wide-ranging implications for progress in psychological science and the latter's successful application to practical problems.

Authors:
Peder Mortvedt Isager [1]

[1] Eindhoven University of Technology

Title:
Quantifying Replication Value

Session & Time:
"Selecting, Designing and Analyzing Replication Studies" - March 13, 11:00am - 11:30am

Abstract:
Background
The concept of replication is a central value of empirical science. At the same time scientists do not regard every replication as equally valuable. Even though replications are a cornerstone of empirical science (Bertamini & Munafò, 2012; Falk, 1998; Jasny, Chin, Chong, & Vignieri, 2011; Koole & Lakens, 2012; Moonesinghe, Khoury, & Janssens, 2007; Rosenthal, 1990; Schmidt, 2009), most researchers will agree that conducting 20 direct replications of the classic and extremely robust Stroop color-naming task (Stroop, 1935) would not be the best way to spend one's grant money.
This raises an important question: when is a replication of an empirical finding of sufficient valuable to the scientific community that it should be be performed? Given limited resources, one could also ask: which among currently published findings are the most valuable to replicate? Some discussion of the circumstances under which replication efforts are more or less beneficial has already occurred in the wake of increased replication efforts in psychology (Brandt et al., 2014; Coles, Tiokhin, Scheel, Isager, & Lakens, 2018), and recently suggestions have been put forward for how to select target studies for replication (Field, Hoekstra, Bringmann, & van Ravenzwaaij, 2018; Kuehberger & Schulte-Mecklenbeck, 2018). A comprehensive evaluation of the factors that could be used to quantify the replication value of a study is currently lacking, which is becoming increasingly important now that more replication studies are funded, performed, and published.

Objectives
We propose a quantitative approach to help researchers, editors and funders evaluate and compare the replication value of original findings. Our approach rests on two fundamental assumptions: (1) That close replication (LeBel, Berger, Campbell, & Loving, 2017; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018) is in principle a worthwhile endeavor, and (2) that there are more original observations worth replicating than we currently have the resources to replicate. In order to help researchers determine which among many findings might be the most promising candidates for replication, we outline a formula-based approach that can be relatively easily used to quantify the replication value of original findings.
We propose that two components determine the replication value of empirical findings: (i) the *impact* of the effect, and (ii) the *corroboration* of the effect. Impact indicates the influence that the effect has had on scientific theory, research activities, or in society. All else being equal, findings that have had more impact are more important to replicate than findings that have had less impact. Corroboration indicates the empirical observations bearing on the finding,

as well as the quality of these observations. As the corroboration of a particular finding increase, it becomes less important to replicate this finding, relative to a finding with little corroboration.
The purpose of a replication value formula is to clarify how one intends to weigh the factors one considers important against one another, and to standardize parts of the study selection procedure. Because these formulas can be calculated quickly (and sometimes even automatically), they can be powerful tools for exploring a large set of studies to discover original findings that are particularly replication-worthy, assuming that the input to the formula is meaningful. Their ultimate goal is to make sure that resources spent on replication are efficiently utilized and that all relevant options for study choice can be considered when a replication effort is initialized.

Research Questions

1) What factors are considered important for determining the replication value of a particular finding or result?
2) Can we create a formula that is able to yield meaningful quantitative estimates of the relative replication value of empirical findings, based on metrics related to the impact and the corroboration of the finding?

Approach & Preliminary results

To assess whether our conceptualization of replication value is in line with evaluations of replication value in the broader community of researchers, and to better understand how replicating researchers justify decisions of study choice, we conducted a literature review of justifications of study selection in 85 replication reports.
The literature review suggests that researchers use many different information sources to assess replication value that could be subsumed under the categories of impact and corroboration (e.g. citation impact, theoretical importance, imprecise estimates, lack of prior replication). However, it is also clear that some types of information cannot easily be quantified (e.g. theoretical importance), and it is clear that factors other than the value of replication matter for the evaluation process as well (e.g. feasibility).
We are currently in the process of constructing one version of a replication value formula that captures the *impact* and *corroboration* of a finding. Once a formula has been derived, we aim to evaluate whether the candidate studies returned by the formula track researchers' qualitative judgements of relative replication value. We will pursue this through two lines of inquiry. First, we will calculate replication value for a large number of studies in the psychological literature and evaluate the face-validity of the recommendations produced, as well as inspect formula recommendations for examples where the true replication is known to be very high or very low (e.g. Stroop, 1935). Second, we will design a study to assess whether formula-based replication value estimate is able to predict researchers' intuitive evaluation of replication value.
Preliminary assessment of face-validity for data in the Curate Science database suggests that the formula yields sensible estimates of replication value for cases where true replication value is known. At the time of the conference, we expect to have completed a comprehensive evaluation of formula performance for both the Curate Science database and a large sample of published studies from the psychological literature. In addition, we will be

able to present the planned experimental design for the study that will compare formula recommendations to researchers' intuitive judgements.

References

Bertamini, M., & Munafò, M. R. (2012). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science*, *7*(1), 67–71. https://doi.org/10.1177/1745691611429353

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Coles, N., Tiokhin, L., Scheel, A., Isager, P., & Lakens, D. (2018). The Costs and Benefits of Replication Studies. https://doi.org/10.17605/osf.io/c8akj

Falk, R. (1998). Replication-A Step in the Right Direction: Commentary on Sohn. *Theory & Psychology*, *8*(3), 313–321. https://doi.org/10.1177/0959354398083002

Field, S., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2018). When and Why to Replicate: As Easy as 1, 2, 3? *Open Science Framework*. https://doi.org/10.17605/osf.io/3rf8b

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and Again, and Again ... *Science*, *334*(6060), 1225–1225. https://doi.org/10.1126/science.334.6060.1225

Koole, S. L., & Lakens, D. (2012). Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science*, *7*(6), 608–614. https://doi.org/10.1177/1745691612462586

Kuehberger, A., & Schulte-Mecklenbeck, M. (2018). Selecting target papers for replication. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000742

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most Published Research Findings Are False—But a Little Replication Goes a Long Way. *PLoS Medicine*, *4*(2), e28. https://doi.org/10.1371/journal.pmed.0040028

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior & Personality*, *5*(4), 1–30.

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. https://doi.org/10.1037/a0015108

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Authors:
Marc Jekel [1], Andreas Glöckner [1,2], Susann Fiedler [2,3], Ramona Allstadt Torras [3,4], Angela Dorrough [1], Dorothee Mischkowski [1], Nicole Franke [3], Janik Goltermann [5], Stefanie Miketta [3]

[1] University of Cologne; [2] MPI Collective Goods Bonn; [3] University of Hagen; [4] University of Bonn; [5] University of Münster

Title:
The Impact of Open Science Practices on Research Methodology: A Case Study for Research in Judgment and Decision Making

Session & Time:
"Open Science and its Impact on Scientific Practice" - March 12, 4:00pm - 4:30pm

Abstract:
In 2011, the journal Judgment and Decision Making introduced as one of the first journals in psychology a standard request for data at submission of an article, which are also used for checks in the review process. In a sample of 71 articles published in this journal we investigate the effectiveness of this policy with respect to the prevalence of direct and indirect open science measures and their development between 2012 and 2018. For 100% of the articles data was available, 80% of the original authors responded positively to student requests for cooperation in conducting replications of their publications, for 94% of the articles materials were available or shared by the authors on request, and 96% of the original analyses were reproducible also by students, 30% of them with minor deviations or after further clarification. The usage of a priori power analyses (10%) and the reporting of effect sizes (66%) were considerably lower but increased over time. For only 4% of the articles analysis scripts were directly available, and none of the studies pre-registered hypotheses or used a pre-registered report format. There was no indication of the usage of small underpowered studies (Md(N) = 193, average effect size r = 0.30) and the p-values showed the expected right-skewed distribution without a bunching of p-values just below p = .05. Overall, adoption rates of open science principle are higher than in other fields and the journals policies were successful in fostering adherence to open science principles. A general culture of openness to reproducing analyses and replication of findings was established, which allows for a cumulative development of knowledge to which also student research can contribute.

Authors:
Anand Krishna [1], Sebastian M. Peter [1]

[1] Julius-Maximilians-Universität Würzburg

Title:
Estimating the prevalence and antecedents of questionable research practices in student theses in psychology from self-reports

Session & Time:
"Academic Practices in Competitive Normal Science" - March 14, 3:30pm - 4:00pm

Abstract:
Background
An important instigator of the current movement towards open science was the realization in the field that exploiting researcher degrees of freedom could inflate false-positive results to astounding degrees (Simmons et al., 2011). Combined with several high-profile cases of data fraud, research investigating the prevalence of these questionable research practices (QRPs; e.g. Agnoli et al., 2017; Fiedler & Schwarz, 2016; John et al., 2012) was received with broad interest and spurred endeavors to reduce their impact in psychological science (Open Science Framework; Grahe et al., 2018). However, no work has yet focused on the prevalence of QRPs in student work even though this is an important index of both the effectiveness of education procedures and the future of the field. Furthermore, the interaction of supervisor and student in thesis work offer an avenue to investigating to what extent QRP use is determined by environmental factors, such as the perception of what authority figures think of QRPs and the freedom to influence certain aspects of the study.

Objectives
This survey aimed to investigate both the prevalence of QRP use in student theses and its antecedents based on students' self-reports. The predictors assessed were attitudes towards QRPs, stress, motivation to write the thesis, perceived supervisor attitudes, and beliefs that good science leads to significant results and that significant results lead to better grades.

Hypotheses
Positive attitudes towards QRPs and stress should increase reported QRP use, higher motivation to complete the thesis should reduce it. Perceived supervisor attitudes towards QRPs and beliefs that good science leads to significant results as well as that significant results lead to better grades should increase reported QRP use both directly and indirectly via attitudes towards QRPs. All of the predicted relationships should be stronger for reporting and analysis QRPs than for study design QRPs due to students' greater freedom in implementing QRPs in reporting and analysis.
Method: 207 German university students either currently working on their thesis or having finished it less than two years prior to data collection responded to an online survey in 2016 that included the target variables as well as further questionnaires on other aspects of the

thesis. All relevant variables were assessed via self-report. Statistical regression and mediation analyses were conducted using the PROCESS macro (Hayes, 2017).

Results

Self-reported QRP prevalence was comparable to that estimated in previous studies (e.g. Fiedler & Schwarz, 2016). However, the majority of students reported engaging in one or less QRPs in their thesis, with more than 88% reporting engaging in two or less. In general, students showed moderately negative attitudes towards QRPs. Although stress did not predict reported QRP use, students' motivation was related to lower reported QRP use and positive QRP attitudes were related to higher reported QRP use for reporting and analysis QRPs. Beliefs that good science leads to significant results and that significant results lead to good grades exerted neither a direct nor an indirect effect on reported QRP use, but perceived supervisor attitudes did exert a direct effect on all reported QRP use and an indirect effect on reported reporting and analysis QRP use. In general, relationships were stronger for reported and analysis QRPs, as predicted.

Conclusions

Although the field of psychology had been discussing the issue of QRP use for several years prior to data collection, our data are sobering: students reported engaging in similar amounts of QRPs to career academics. This engagement was at least partially predicted by their attitudes towards QRPs, which in turn were affected by their perception of what their supervisor thinks. These findings underline the importance of communicating clear norms about QRP use to students when they engage in their practical thesis work. Although students do not generally endorse QRPs, presumably due to their methodological training, they may still be influenced by their supervisors if these supervisors are perceived as endorsing QRPs. We could find no evidence that beliefs about the necessity of significant results for good science or good grades impacted QRP attitudes or reported QRP use, but the data did not allow for a strong conclusion that there is no relationship between these variables, as the initial variance of these predictors was low due to a majority of participants strongly disagreeing. Finally, motivation appears to be a protective factor against QRP use that is possible to influence for supervisors. Although alternative causal models may obtain due to the cross-sectional nature of these data, the conclusion that projecting a critical attitude towards QRPs and motivating students can reduce the likelihood of them implementing QRPs is plausible. Furthermore, the differences between study design QRPs and those of reporting and analysis in our data shed some light on how structural constraints can influence whether an individual's propensity to engage in QRPs will affect their behavior. Beyond informing teaching practice with undergraduate students, these data provide some indications of factors that can prevent QRP use with junior researchers, in particular with regard to their relationship with their supervisors.

Authors:
Audrey Linden [1], Johannes Hönekopp [1]

[1] Northumbria University

Title:
Heterogeneity in the Results of Close and Conceptual Replications: Implications for Scientific Progress and Practical Applications

Session & Time:
"Assessment of Heterogeneity in Meta-Analysis" - March 13, 5:30pm - 6:00pm

Abstract:
Background
Replicability of the results of original studies and close replications have become a prominent concern in psychology. In particular, heterogeneity, i.e. variability in the results of replication studies over and above what is expected due to sampling error, has received increased attention, because it decreases the statistical power of studies. Heterogeneity-induced low power might therefore explain the low success in Open Science Collaboration's (2015) 100 close replications of studies in cognitive and social psychology (Stanley, Carter, & Doucouliagos, in press).

Objectives
To derive a reliable estimate of heterogeneity in close replication studies and to compare this to levels of heterogeneity in meta-analyses (conceptual replications). To investigate to what extent moderators explain heterogeneity. To investigate some of the possible causes of high heterogeneity.

Research question(s) and/or hypothesis/es
We expect heterogeneity in close replications to be low. Based on the results in Open Science Collaboration (2015), we expect that heterogeneity in conceptual replications will be larger in social psychology than in cognitive psychology. Based on Michell et al. (2012) we expect that heterogeneity in social psychology will be larger than in organisational psychology.

Method/Approach
We analysed heterogeneity in all available Many-Labs type replications (n = 40). We compared this against heterogeneity in conceptual replications observed in 147 meta-analyses in cognitive, social, and organisational psychology. In all cases, we used Cohen's d as a measure of effect size, and τ, the standard deviation of true population effect sizes, to quantify heterogeneity.

Results/Findings
Heterogeneity in close replications was found to be typically very low (T = 0.08). In contrast, heterogeneity turned out to be very high in the average (conceptual replication) meta-analysis (T = 0.33). This was largely unexplained by moderators, but appeared

strongly driven by the magnitude of effects. The distinctive success rates of close replications in cognitive and social psychology (Open Science Collaboration) were not reflected in those disciplines' heterogeneity levels in conceptual replications.

Conclusions and implications (expected)
We argue (contrary to Stanley et al., in press) that this pattern of results indicates sufficient power in Open Science Collaboration's close replications.

Authors:
Helen Niemeyer [1], Robbie C.M. van Aert [2], Sebastian Schmid [3], Dominik Uelsmann [3], Christine Knaevelsrud [1], Olaf Schulte-Herbrueggen [3]

[1] Freie Universität Berlin; [2] Tilburg University; [3] Charité - Universitätsmedizin Berlin

Title:
Comparison of the performance of methods to assess publication bias in real data

Session & Time:
"Assessment of Publication Bias" - March 14, 12:00pm - 12:30pm

Abstract:
Background
Publication bias is widespread within many scientific disciplines and also within psychology. If only published studies are included in a meta-analysis of psychotherapy research, the efficacy of interventions may be overestimated. However, the treatments in evidence-based psychotherapy are mainly selected based on published rather than unpublished research. The presence and impact of publication bias in psychotherapy research remains largely unknown.
Posttraumatic stress disorder (PTSD) is a highly distressing and common condition, and various forms of psychological interventions for treating PTSD have been investigated in a large number of studies. A comprehensive statistical assessment of publication bias in meta-analyses of psychotherapeutic treatments for PTSD has not been conducted, even though a considerable number of statistical methods to investigate the presence and impact of publication bias have been developed in recent years.

Objectives
We compare the performance of six state-of-the-art publication bias methods on a large-scale data set by re-analyzing all meta-analyses which investigate the efficacy of psychotherapeutic interventions for PTSD.

Research question
We aim at investigating the amount of publication bias in all meta-analyses on the efficacy of psychotherapeutic treatment for PTSD and to compare the performance of methods to assess publication bias. A comparison on real data is not as straightforward as in simulation studies, since the true amount of publication bias is unknown in real data. Hence, the performance of the methods will be examined by comparing them to the other included methods.

Method
We screened the databases PsycINFO, Psyndex, PubMed, and the Cochrane Database for all published and unpublished meta-analyses in English or German up to 5th September 2015. In addition, a snowball search system was used. Meta-analyses were required to meet the following inclusion criteria: 1) A psychotherapeutic intervention was evaluated. 2) The

intervention aimed at reducing subclinical or clinical PTSD. 3) A summary effect size was provided.
We included only data sets where the null hypothesis of homogeneous true effect size was not rejected, because the statistical methods become biased if the true effect sizes are heterogeneous. This hypothesis was tested by means of the Q-test and quantified by I^2. Moreover, we excluded all data sets of a meta-analysis that included five or fewer trials, as the methods to detect publication bias are underpowered if the number of studies is too small.
We included the following methods to test whether publication bias was present in a meta-analysis: Egger's regression test, rank-correlation test, TES, and p-uniform's publication bias test. Four different methods were included to estimate the effect size and test the null hypothesis of no effect: traditional meta-analysis, trim and fill, PET-PEESE, and p-uniform. The degree of agreement among the methods was examined by means of Loevinger's H, because the publication bias tests and tests of the null hypothesis of no effect resulted in a dichotomous decision (statistically significant or not).

Results
The literature search resulted in 7,647 hits including duplicates. The screening process reduced this number to 502 meta-analyses, of which 83 dealt with the efficacy of psychotherapeutic interventions for PTSD and were included. The meta-analyses included a total number of 2,131 data sets, of which 93 data sets from 24 meta-analyses fulfilled all inclusion criteria. They included a median number of 7 studies. Since publication bias tests have low statistical power if the number of effect sizes is small in a meta-analysis, the characteristics of many of the data sets are not well-suited for methods of detecting publication bias. The median number of statistically significant effect sizes in the data sets was 3.
Of all methods Egger's regression test detected publication bias the most, i.e. in 17 data sets (18.3%). At most two methods detected publication bias test in the same data set, which occurred in 4 data sets (4.3%). Loevinger's H varied between -.075 and 1. For the test of no effect, Loevingers´s H varied between .668 and 1. When estimating effect sizes corrected for publication bias, results show that especially estimates of PET-PEESE were closer to zero than traditional meta-analysis and that the standard deviation of the estimates of PET-PEESE and p-uniform was larger than traditional meta-analysis and trim and fill. The mean of the difference in effect size estimate between PET-PEESE and the traditional meta-analytic estimate was -0.108 (SD = 0.886). P-uniform was applied to a subset of 72 data sets, because at least one study in a data set has to be statistically significant for this method. The mean of the difference in effect size estimate of p-uniform and traditional meta-analysis was 0.002 (SD = 0.355). Estimates of PET-PEESE were especially unrealistic if there was a small number of effect sizes in a data set in combination with small variation in the standard errors of the primary studies. P-uniform's estimates were unrealistically large or small when a small number of statistically significant effect sizes were observed with p-values just below the α-level.

Conclusions
Our study is the first to apply a multitude of publication bias methods to a large-scale real data set. Publication bias tests did not result in the same conclusion in the majority of the

data sets which is unlikely if extreme publication bias was present. No clear indications for overestimated effect sizes were observed when comparing effect size estimates of traditional meta-analysis with methods to correct for publication bias. However, the assessments of publication bias in psychotherapy research may have lacked statistical power to detect publication bias. Moreover, the conclusion regarding statistical significance of the test of no effect often changed when correcting for publication bias with PET-PEESE and p-uniform compared to traditional meta-analysis. This is at least partly caused by the less precise effect size estimates of these methods since the effect size estimates corrected for publication bias did not provide strong evidence for overestimation caused by publication bias. Future research is needed to study the convergence and divergence of publication bias tests as a function of publication bias and the number of primary studies in a meta-analysis.

Authors:
Anton Olsson-Collentine [1], Jelte Wicherts [1], Marcel A.L.M. van Assen [1,2]

[1] Tilburg University; [2] Utrecht University

Title:
Heterogeneity in direct replications in psychology and its association with effect size

Session & Time:
"Assessment of Heterogeneity in Meta-Analysis" - March 13, 6:00pm - 6:30pm

Abstract:
In meta-analysis, the heterogeneity of an effect size (henceforth referred to as heterogeneity) is a measure of an effect's susceptibility to changes in four contextual factors; the 1) sample population, 2) settings, 3) treatment variables and 4) measurement variables (e.g., Campbell & Stanley, 2015). Heterogeneity is of concern for several reasons. First, heterogeneity can have important practical consequences. For example, mental health interventions that work for some patients but not for others (or even have negative consequences). Second, unaccounted for heterogeneity suggests a theory is unable to predict all contextual factors of importance to its claims. Third, the possibility of heterogeneity can create controversy in the interpretation of replication results. An explanation often suggested when a study 'fails' to replicate is that the studied effect is more heterogeneous than (perhaps implicitly) claimed originally.

We examined the evidence for heterogeneity and explored the association between heterogeneity and effect size in a sample of 37 effect sizes from ten pre-registered multi-lab direct replication projects in psychology. The 37 effect sizes represent the primary effects from all multi-lab replication projects available at the time of data collection (2018/02/01 – 2018/03/31) according to http://curatescience.org. To better interpret the heterogeneity estimates we also estimated power of each project to find zero/small/medium/large heterogeneity using simulations. Our analyses provide information on how two contextual factors (sample population and settings) may affect consistency or heterogeneity of effects in direct replications in psychology, and on the precision of its estimate.

We found limited heterogeneity; only 7/37 (19%) effects had significant heterogeneity, and most effects (32/37; 86%) were most likely to have zero to small heterogeneity. Power to detect small heterogeneity was low for all projects (mean 36%), but good to excellent for medium and large heterogeneity. Our findings thus show little evidence of widespread heterogeneity in direct replication studies in psychology, implying that citing heterogeneity as a reason for non-replication of an effect is unwarranted unless predicted a priori. We also found a strong correlation between observed effect size and heterogeneity in our sample (r = .78), suggesting that heterogeneity and moderation of effects is implausible for a zero average true effect size, but increasingly plausible for larger average true effect size.

Authors:
Frank Renkewitz [1], Melanie Keiner [1]

[1] University of Erfurt

Title:
How to detect publication bias in psychological research? A comparative evaluation of six statistical methods

Session & Time:
"Assessment of Publication Bias" - March 14, 11:30am - 12:00pm

Abstract:
Publication biases and questionable research practices are assumed to be two of the main causes of low replication rates observed in the social sciences. Both of these problems do not only increase the proportion of false positives in the literature but can also lead to severely inflated effect size estimates in meta-analyses. Methodologists have proposed a number of statistical tools to detect and correct such bias in meta-analytic results. We present an evaluation of the performance of six of these tools in detecting bias. To assess the Type I error rate and the statistical power of these tools we simulated a large variety of literatures that differed with regard to underlying true effect size, heterogeneity, number of available primary studies and variation of sample sizes in these primary studies. Furthermore, simulated primary studies were subjected to different degrees of publication bias. Our results show that the power of the detection methods follows a complex pattern. Across all simulated conditions, no method consistently outperformed all others. Hence, choosing an optimal method would require knowledge about parameters (e.g., true effect size, heterogeneity) that meta-analysts cannot have. Additionally, all methods performed badly when true effect sizes were heterogeneous or primary studies had a small chance of being published irrespective of their results. This suggests, that in many actual meta-analyses in psychology bias will remain undiscovered no matter which detection method is used.

Authors:
Anne M. Scheel [1]

[1] Eindhoven University of Technology

Title:
Positive result rates in psychology: Registered Reports compared to the conventional literature

Session & Time:
"Open Science and its Impact on Scientific Practice" - March 12, 4:30pm - 5:00pm

Abstract:
Background
Several studies have found the scientific literature in psychology to be characterised by an exceptionally high rate of publications that report 'positive' results (supporting their main research hypothesis) on the one hand, and notoriously low statistical power on the other (Sterling, 1959; Fanelli, 2010; Maxwell, 2004). These findings are at odds with each other and likely reflect a tendency to under-report negative results, through mechanisms such as file-drawering, publication bias, and 'questionable research practices' like p-hacking and HARKing. A strong bias against negative results can lead to an inflated false positive rate and inflated effect sizes in the literature, making it difficult for researchers to build on previous work and increasing the risk of ineffective or harmful 'evidence-based' applications and policies.
In 2013, Registered Reports (RRs) were developed as a new publication format to reduce under-reporting of negative results by mitigating file-drawering, publication bias, and questionable research practices: Before collecting and analysing their data, authors submit a protocol containing their hypotheses and methods to a journal, where it gets reviewed and, if successful, receives 'in-principle acceptance' which guarantees publication once the results are in, regardless of the outcome. Given their bias-reducing safeguards, we should expect a lower positive result rate in RRs compared to the non-RR literature, but to date no structured comparison of RRs and non-RRs has been offered.

Objectives
Fanelli (2010) presented a simple method to assess the positive result rate in a large sample of publications. We used his method to replicate his results for the (non-RR) psychology literature since 2013 and compare it to all published RRs in psychology.

Hypothesis
Using Fanelli's method, we tested the hypothesis that published RRs in psychology have a lower positive result rate than non-RRs in psychology published in the same time range (2013-2018).
We would reject this hypothesis if the difference between RRs and non-RRs were found to be significantly smaller than 6%.

Method

To obtain the non-RR sample, we applied Fanelli's (2010) sampling strategy: We searched all journals listed in the 'Psychiatry/Psychology' category of the Essential Science Indicators database for the phrase 'test* the hypothes*' and picked a random sample of 150 publications of all search results, deviating from Fanelli only in restricting the year of publication to 2013-2018.

To obtain the RR sample, we relied on a list of published RRs curated by the Center for Open Science (https://www.zotero.org/groups/479248/osf/items/collectionKey/KEJP68G9?) which at the time had 152 entries, and excluded all publications that were not in psychology or not certainly RRs, leaving 81 publications.

The positive result rate was determined by identifying the first hypothesis mentioned in the abstract or full text and coding whether it was (fully or partially) supported or not supported, and then for each group calculating the proportion of papers that reported support.

Methods and analyses were preregistered at https://osf.io/s8e97/.

Results

Eight non-RRs and 13 RRs were excluded because they either did not test a hypothesis or could not be coded for other reasons, leaving 142 non-RRs and 68 RRs. The positive result rate was 95.77% for non-RRs and 42.65% for RRs. The proportion difference was significantly different from zero (one-sided Fisher's exact test, alpha = .05), $p < .0001$, and not significantly smaller than our smallest effect size of interest of 6% in an equivalence test, $Z = -7.564$, $p > .999$.

For an exploratory analysis we also coded whether or not a paper contained a replication of previous work and found that none of the non-RRs, but two thirds (42/68) of the RRs did. The positive result rate for replication RRs was slightly lower (35.71%) than for original RRs (53.85%), but this difference was not significant, $p = .112$.

Conclusions and implications

In 2010, Fanelli reported a positive result rate of 91.5% for the field of psychology. Using the same method, we found a rate of 95.77% for the time between 2013 and 2018, suggesting that the rate has not gone down in recent years. In contrast, with only 42.65% the new population of Registered Reports shows a strikingly lower positive result rate than the non-RR literature. This difference may be somewhat smaller when focussing only on original work, but the RR population is currently too small to draw strong conclusions about any differences between replication and original studies.

Our conclusions are limited by the different sampling procedures for RRs and non-RRs and by the observational nature of our study, which did not allow us to account for potential confounding factors. Nonetheless, our results are in line with the assumption that RRs reduce under-reporting of negative results and provide a first estimate for the difference between this new population of studies and the conventional literature.

References

Fanelli, D. (2010). 'Positive' Results Increase Down the Hierarchy of the Sciences. PLOS ONE, 5(4), e10068. doi: 10.1371/journal.pone.0010068

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. Psychological Methods, 9(2), 147–163. doi: 10.1037/1082-989X.9.2.147
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. doi: 10.1080/01621459.1959.10501497

Authors:
Peter Steiner [1], Vivian Wong [2]

[1] University of Wisconsin-Madison; [2] University of Virginia - Charlottesville

Title:
Assessing the Correspondence of Results in Replication Studies

Session & Time:
"Selecting, Designing and Analyzing Replication Studies" - March 13, 12:00pm - 12:30pm

Abstract:
Background
Reproducibility is a hallmark of science. Instead of relying on causal conclusions from a single experimental or non-experimental study, scientific knowledge is best achieved through careful replication of studies, or meta-analysis of results from multiple studies. In replication studies, researchers assess whether the original study results are reproduced by looking at the direction and magnitude of effects, as well as statistical significance patterns of results. Researchers may also conduct direct tests of statistical difference between the effect estimates of the original and replication study. Steiner and Wong (2018) distinguish between two classes of measures for assessing correspondence in results. The first consist of distance-based measures, which estimates the difference in the original and replicated effect. The second class of metrics is what Steiner and Wong (2018) call "conclusion-based measures." These approaches assess correspondence in study results by looking at the size, direction, and statistical significance patterns of results.

Objective/Research Question
This paper focuses on distance-based significance tests for assessing whether the effect estimates of the original and replication study actually replicate within the boundaries of statistical uncertainty. All distance-based tests use the difference in the effect estimates as the starting point. However, they differ with respect to the null- and alternative hypotheses under investigation. In a comparative simulation study, we assess the properties of three different significance tests for probing the difference or equivalence of two effect estimates. The first test is the standard null hypothesis significance test to which we refer as difference test because the alternative hypothesis claims a difference in effects. Thus, the difference test rejects the null hypothesis of no difference if the p-value is sufficiently low. The second test is the less well-known equivalence test where the alternative hypothesis claims the equivalence of the two effects while the null hypothesis states a difference in effects (Tryon, 2001). This test protects against the common type II error of difference tests, that is, the failure to reject the false null hypothesis of no difference. However, an underpowered equivalence test might fail to reject the false null hypothesis of a difference in effects. We suggest the use of a combined difference and equivalence test, the correspondence test (Steiner & Wong, 2018; Tryon & Lewis, 2008). The correspondence test has four possible outcomes: equivalence, difference, trivial difference, and indeterminacy. Its advantage is that it indicates indeterminacy whenever the two studies lack sufficient power to demonstrate either a significant difference or equivalence. Many of the replication efforts undertaken so

far would presumably obtain an indeterminate outcome from the correspondence test. Thus, the objective of the paper is to compare the three tests for assessing correspondence in replication results under different scenarios.

Research Question
(1) In which scenarios do difference tests (fail to) perform well?
(2) In which scenarios do equivalence tests show excellent or poor performance?
(3) Does the combined correspondence test outperform the difference and equivalence test?
(4) What are the power requirements for the original and replication study to guarantee conclusive correspondence test results?

Method/Approach
First, we use statistical arguments to highlight the strengths and weaknesses of each test from a theoretical point of view. Second, we use a simulation study to compare the performance of the three distance-based tests under different scenarios. The scenarios are defined with respect to variations in (a) each study's minimum detectable effect size (i.e., sample size and magnitude of the underlying true effect, which directly relates to the studies' power) and (b) the difference in the true effects (including the equivalence of effects). True effect differences might result from effect heterogeneities across sites, populations, or settings, or from biases due to imperfect implementation of at least one of the studies.

Results/Findings
Theory and our simulations suggest that the difference test regularly fails to indicate a true difference in effects whenever one of the two studies is insufficiently powered. That is, the probability of a type II error is high. A similar issue occurs with the equivalence test which fails to indicate equivalence (i.e., reject the null hypothesis of a difference) if one of the studies is insufficiently powered. Importantly, failing to reject the null hypothesis in an equivalence test does not imply that the effects actually differ. Here, the correspondence test has a clear advantage because if equivalence cannot be established, the correspondence test is able to distinguish between insufficient power (indeterminacy) and a significant difference. However, if researchers want to avoid an indeterminate outcome from the correspondence test both studies need to be sufficiently powered.
The simulation results also indicate that both studies' power must be significantly larger than what researchers usually plan for in a single study. This is so because the comparison of two effect estimates from independent studies requires that both effects are estimated with high efficiency.

Conclusion & Implications
Theoretical considerations and the simulation results suggest that researchers interested in replication should use the correspondence test for assessing whether the effects of an original and replication study successfully reproduce or fail to reproduce. A major advantage of the correspondence test is that it indicates an indeterminate test outcome if the studies lack sufficient power to establish either a significant difference or equivalence. It becomes also clear that testing the correspondence of replication efforts requires highly powered studies.

Authors:
Isabel Thielmann [1]

[1] University of Koblenz-Landau

Title:
Balancing errors as an approach towards better use of larger samples in psychological research

Session & Time:
"Consequences of Publication Bias on Statistical Evidence Generation" - March 13, 3:30pm - 4:00pm

Abstract:
A well-documented concern about psychological research that has been frequently stressed during the current replicability crisis maintains that studies are often based on too small samples and tests thus yield insufficient statistical power. Correspondingly, it has been repeatedly emphasized that a vital step in overcoming the low replicability of psychological studies is to recruit larger samples whenever possible. However, despite the indubitable importance of larger samples – and the value of corresponding policy changes by editors and reviewers – increasing power will, all else being equal, only reduce one type of error (namely, β) whereas α is held constant at 5%. As a consequence, errors may be severely imbalanced, which is problematic for at least two reasons: First, retaining imbalanced errors implicitly assigns greater importance to one error over the other by affecting the "relative seriousness" of errors. In the extreme, increasing sample sizes and thus statistical power will inadvertently assign greater seriousness to β than to α if the latter is held constant at .05. Second, and more importantly, fixing α at .05 ultimately means that the statistical test cannot achieve consistency, meaning that tests will not point to the true state of the world even in the large sample limit. By implication, the conclusiveness of (non-)significant results will remain limited despite large samples and high power. To demonstrate this, we conducted two simulations comparing the Positive Predictive Value (PPV) and the proportion of correct inferences implied by fixed α versus balanced errors (i.e., α = β). For PPV, simulations showed that once the sample size is sufficiently large to render β < α (i.e., 1–β > .95), adjusting α corresponding to β results in higher PPV than holding α fixed at .05, irrespective of the probability *p*(H1) that the alternative hypothesis is true. For the proportion of correct inferences, in turn, results imply that balanced errors are to be preferred over fixed α in two situations: (1) whenever β < α (i.e., as soon as the sample size yields β < .05) which holds practically independent of *p*(H1) and (2) whenever *p*(H1) > .50, practically irrespective of the absolute magnitude of α and β. Fixing α = .05, by contrast, is only superior whenever β > α and *p*(H1) < .50, that is, whenever statistical power is not entirely satisfactory and the alternative hypothesis is known to be less likely to hold than the null. Overall, we therefore advocate for extending the calls for higher statistical power by also calling for balanced errors based on straightforward compromise power analyses if samples are large. In other words, to fully exploit the advantages of large samples and to render statistical tests consistent, researchers should not blindly replace a general lack of power with increasingly imbalanced errors, but instead strive for smaller error probabilities *in general*.

Authors:
Leonid Tiokhin [1], Maxime Derex [2]

[1] Eindhoven University of Technology; [2] University of Exeter

Title:
Competition for novelty in an information-sampling game

Session & Time:
"Academic Practices in Competitive Normal Science" - March 14, 4:00pm - 4:30pm

Abstract:
*This abstract is adapted from a registered report accepted in-principle at Royal Society Open Science.

Many factors plausibly affect the reliability of science (Munafò et al., 2017). At the heart of these are incentive structures: by determining the professional payoffs for various types of research, incentives shape scientists' research decisions (Nosek, Spies, & Motyl, 2012). One longstanding incentive in academic science is rewarding priority of discovery. Over 50 years ago, sociologist of science Robert Merton noted how this norm might benefit science: rewarding priority can incentivize scientists to invest effort to quickly solve important problems and share their discoveries with the scientific community (Merton, 1957). Nonetheless, scholars have also had longstanding concerns about the repercussions of this norm. Charles Darwin thought that rewarding priority by naming species after their first-describers incentivized biologists to produce "hasty and careless work" by "miserably describing a species in two or three lines." ((Merton, 1957), p. 644). More recently, concerns over the consequences of rewarding priority have led the academic journals eLife and PLOS Biology to offer "scoop protection" (i.e. allowing researchers to publish findings identical to those already published in the same journal) in attempts to reduce the disproportionate payoffs to scientists who publish first (Marder, 2017; The PLOS Biology Staff Editors, 2018; Yong, 2018). In the editorial justifying their new policy, The PLOS Biology Staff Editors write "…many of us know researchers who have rushed a study into publication before doing all the necessary controls because they were afraid of being scooped. Of course, healthy competition can be good for science, but the pressure to be first is often deleterious…" (The PLOS Biology Staff Editors, 2018).
Despite these reasonable concerns, there is little empirical evidence for the hypothesis that competitive pressures to publish cause individuals to produce lower-quality research. In focus-group discussions with mid and early-career researchers, scientists acknowledge that competition incentivizes them to conduct careless work (Anderson, Ronning, De Vries, & Martinson, 2007), but laboratory experiments investigating competition more broadly demonstrate that competition also promotes individual effort (Baer, Vadera, Leenders, & Oldham, 2013; Balietti, Goldstone, & Helbing, 2016; Dechenaux, Kovenock, & Sheremeta, 2015; Dohmen & Falk, 2011; Gneezy, Niederle, & Rustichini, 2003; Niederle & Vesterlund, 2007). As a consequence, it is unclear how competition in general, and competition for priority in particular, affects research quality. On the one hand, competition might cause

researchers to make dubious claims based on inadequate data. On the other, competition might encourage researchers to gather data more efficiently.

Given the difficulty of experimentally manipulating incentives in real-world scientific practice, we develop a simple game that mimics aspects of scientific investigation. In our experiment, individuals gather data in order to guess true states of the world and face a tradeoff between guessing quickly and increasing accuracy by acquiring more information. To test whether competition affects accuracy, we compare a treatment in which individuals are rewarded for each correct guess to a treatment where individuals face the possibility of being "scooped" by a competitor. In a second set of conditions, we make information acquisition contingent on solving arithmetic problems to test whether competition increases individual effort (i.e. arithmetic-problem solving speed). We find that competition causes individuals to make guesses using less information, thereby reducing their accuracy. We find no evidence that competition increases individual effort. Our experiment provides proof of concept that rewarding priority of publication can incentivize individuals to acquire less information, producing lower-quality research as a consequence. More generally, it provides one example of the type of empirical work that is necessary to move beyond verbal arguments about the effects of incentive structures on scientists' behavior.

References

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*(4), 437–461.

Baer, M., Vadera, A. K., Leenders, R. T., & Oldham, G. R. (2013). Intergroup competition as a double-edged sword: How sex composition regulates the effects of competition on group creativity. *Organization Science*, *25*(3), 892–908.

Balietti, S., Goldstone, R. L., & Helbing, D. (2016). Peer review and competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences*, 201603723.

Dechenaux, E., Kovenock, D., & Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, *18*(4), 609–669. https://doi.org/10.1007/s10683-014-9421-0

Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, *101*(2), 556–90.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Marder, E. (2017). Scientific Publishing: Beyond scoops to best practices. *ELife*, *6*, e30076. https://doi.org/10.7554/eLife.30076

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review*, *22*(6), 635–659.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., … Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/doi:10.1038/s41562-016-0021

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

The PLOS Biology Staff Editors. (2018). *The importance of being second*. Public Library of Science San Francisco, CA USA.
Yong, E. (2018, February 1). In Science, There Should Be a Prize for Second Place. *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2018/02/in-science-there-should-be-a-prize-for-second-place/552131/

Authors:
Olmo van den Akker [1]

[1] Tilburg University

Title:
How Do Academics Assess the Results of Multiple Experiments?

Session & Time:
"Academic Practices in Competitive Normal Science" - March 14, 4:30pm - 5:00pm

Abstract:
Introduction
In both social and experimental psychology a single study is typically not considered to be sufficient to test a theory, and multiple study papers are the norm. In this project, we consider how researchers assess the validity of a theory when they are presented with the results of multiple studies that all test that theory. More specifically, we consider what researchers' beliefs in the theory are as a function of the number of significant vs. nonsignificant studies, and whether this relationship depends on the type of study (direct or conceptual replication) and the role of the respondent (researcher or reviewer). This information is especially relevant in the context of the current replication crisis in psychology, which has prompted a discussion on what evidence sufficiently corroborates a phenomenon.
In addition, we carry out a preregistered secondary analysis in which we look at individual researcher data to find out which heuristics researchers use when assessing the outcomes of multiple studies. We lump each researcher into one of six categories: those who use Bayesian inference (i.e. the normative approach using Bayes' rule incorporating information about statistical power and the significance level), those who use deterministic vote counting (i.e. those who believe the theory is true if the proportion of significant results is higher than 0.5, will believe the theory is false when that proportion is lower than 0.5, and have a 50/50 belief if the proportion is precisely 0.5), those who use proportional vote counting (i.e. those who equate their belief in the theory to the proportion of significant results), those who average their prior belief with the proportion of significant results, those with irrational response patterns, and those whose response patterns are inconsistent with any of the previous categories.

Method
Sample
Our sample consisted of 505 participants from social and experimental psychology who commonly conduct (as researchers) or judge (as reviewers or editors) experimental research consisting of multiple studies.

Procedure
Our vignette study involved eight different scenarios, each presenting the results of four experiments. All presented scenarios stated that other researchers had previously published the results of one experiment, A, and found a statistically significant effect in line with a given theory. The vignette then stated that the participant had conducted (in the 'author' version of

the vignette) or were asked to review (in the 'reviewer' version of the vignette) four experiments that replicated the findings of the original study. The first new experiment, A', was a direct replication of the earlier experiment, whereas the other three experiments (B, C, and D) were conceptual replications. All participants were told to imagine that their prior belief in the theory was 50% and that the number of participants, the costs of all experiments, the nominal significance level, and the statistical power in all five experiments (including the original experiment A) were typical for experimental studies in psychology. After each scenario, we presented subjects with several questions, starting with a set of three general questions regarding the theory and following up with a set of questions concerning the participants' behavior either as author or reviewer. First, participants indicated how they assessed their belief in the theory on the basis of the presented evidence by means of a slider bar, with points going from low probability (0%) to high probability (100%) of the theory being correct. Second, we asked participants to indicate whether they thought that the theory was correct or not based on the outcomes in the scenario ('yes' or 'no'). Third, we asked those in the role of author whether they would submit a paper based on at least one of the experiments to a journal, and those in the role of reviewer whether they would recommend such a paper for publication. Fourth, we asked 'authors' whether they would want to conduct an additional conceptual replication, E, given the results of the earlier experiments, and we asked 'reviewers' whether they would recommend the authors to conduct an additional conceptual replication, E.

Results of the main analysis

We found that participants' belief in the theory increases with the number of significant results, and that direct replications were considered to be more important than conceptual replications for participant's beliefs in the underlying theory. We found no difference between authors and reviewers in their propensity to submit or recommend to publish sets of results, but we did find that authors are generally more likely to desire an additional experiment.

Results of the secondary analysis

The results show that only 6 participants out of the 505 used the normative method of Bayesian inference and that the majority of participants use vote counting approaches that tend to undervalue the evidence for the underlying theory if two or more results are statistically significant.

Conclusions and Discussion

The main results of our study are that:

- researchers valued direct replications more than conceptual replications when deciding on the validity of a theory, perhaps not surprising in the light of of the current popularity of large-scale direct replication efforts like the Many Labs Replication Project and the Reproducibility Project: Psychology
- authors and reviewers contribute equally to publication bias, even though previous research mostly pointed to authors not submitting nonsignificant results as the main cause of publication bias.
- researchers make structural errors when assessing scientific papers with multiple outcomes. Most notable, they use simple heuristics to make sense of this complex situation, which often leads them to undervalue the evidence in favor of a theory. Hopefully, this

information can be used to create methods to educate current and future researchers to avoid making these errors.

Authors:
Martin Voracek [1], Michael Kossmeier [1], Johannes Vilsmeier [1], Rosalie Dittrich [1], Tanja Fritz [1], Caroline Kolmanz [1], Constantin Y Plessen [1], Agnieszka Slowik [1], Ulrich S Tran [1]

[1] University of Vienna

Title:
Long-term trends (1980-2017) in the N-pact factor of journals in personality psychology and individual differences research

Session & Time:
"Metascience" - March 12, 6:00pm - 6:30pm

Abstract:
Recent metascience investigations into the N-pact factor (NF; median sample size of studies published in a journal) have revealed NFs of merely about 100 in fields like social, sport, and exercise psychology. Journal NF has also been shown to correlate negatively with journal impact factors (JIF), implying that the smallest studies appear in the most prestigious journals. In this first long-term and largest NF analysis to date (3699 articles coded), annual NFs of two personality psychology journals were tracked over 38 years since their inception in 1980. Overall NF was about 190, gradually increased over time, and within-journal NF-JIF correlations were positive. Online samples and articles featuring supplemental files presented larger, whereas student samples smaller, NFs than their counterparts. Conspicuous and puzzling distributional irregularities of sample-size numbers included multimodality and surplusses of even-numbered sample sizes and of those just beyond 100. An NF statement, accompanying authors' submitted papers, is suggested.

Authors:
Vivian C. Wong [1], Peter M. Steiner [2]

[1] University of Virginia - Charlottesville; [2] University of Wisconsin-Madison

Title:
Moving from What Works to What Replicates: A New Framework for Evidence Based Policy Analysis

Session & Time:
"Selecting, Designing and Analyzing Replication Studies" - March 13, 11:30am - 12:00pm

Abstract:
Background
Efforts to promote evidence-based practices in medicine and the social sciences (e.g., What Works Clearinghouse) assume that scientific findings are of sufficient validity to warrant their use in decision making. Replication has long been a cornerstone for establishing trustworthy scientific results. At its core is the belief that scientific knowledge should not be based on chance occurrences. Rather, it is established through systematic and transparent methods, results that can be independently replicated, and findings that are generalizable to at least some target population of interest (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015). However, despite consensus on the need to promote replication, there remains considerable disagreement about what constitutes as replication, how a replication study should be implemented, how results from these studies should be interpreted, and whether direct replication of results is even possible (Hansen, 2011).

Objectives
This paper seeks to address these concerns by developing the methodological foundations for a "replication science." The paper introduces the Causal Replication Framework, which defines "replication" as a research design that tests whether two (or more) studies produce the same causal effect within the limits of sampling error. Using potential outcomes notation, the framework provides a clear definition of replication and highlights the conditions under which results are likely to replicate. The paper also demonstrates how different research designs may be used to evaluate the replication of results and identify sources of effect heterogeneity.

Research question(s)
To this end, the paper will address three research questions:
(1) Under the Causal Replication Framework, what is replication?
(2) What research design assumptions are required for successful replication of results?
(3) How may research design features be used to address replication assumptions in high quality, systematic replication studies?

Method/Approach
Over the years, researchers have sought to clarify what is meant by a replication. Most prior definitions have focused on repeating methods and procedures in conducting a replication

study (Schmidt, 2009). Schmidt also describes direct replication as a “methodological tool based on a repetition procedure,” but adds that its purpose is for “establishing a fact, truth or piece of knowledge” (2009, p. 91, emphasis in original).

In the Causal Replication Framework, we begin with the premise that replication is for establishing a “fact, truth or piece of knowledge.” Here, the “piece of knowledge” can be described as the causal effect of a well-defined treatment-control contrast on a clearly specified outcome for a well-defined target population. We refer to this unknown causal effect as the causal estimand, which is the target of inference in the original and replication study. Using a potential outcomes framework, we show that five research design assumptions are required for the direct replication of results:

A1 Treatment & Outcome Stability

A1.1 No hidden variation in treatment and control conditions.

A1.2 No variation in outcome measures.

A1.3 No mode-of-study selection effects.

A1.4 No peer, spillover, or carryover effects.

A2 Equivalence of Causal Estimands

A2.1 Same causal quantity of interest.

A2.2 Identical effect-generating processes.

A2.3 Identical distribution of population characteristics.

A2.4 Identical distribution of setting variables.

A3 Identification of Causal Estimands. In both studies, the causal estimand must be identified using an experimental or quasi-experimental research design.

A4 Unbiased Estimation of Causal Estimands. In both studies, the causal estimand (ATE) is estimable without bias.

A5 Correct Reporting of Estimands, Estimators, and Estimates. For both studies, the estimands, estimators, and estimates are correctly reported.

The replication assumptions highlight the difference between traditional, procedure-based approaches to replication and the Causal Replication Framework. In procedure-based approaches, the goal and purpose of replication is repetition of methods. In the Causal Replication Framework, the goal is for two studies to identify and estimate the same well-defined causal estimand of interest. This means that while two studies may use the same procedures and methods to generate the same corresponding causal effect, it is also possible for two studies to use different methods and procedures as long as they identify and estimate the same well-defined causal estimand of interest.

Results/Findings

Conceptualizing replication through the Causal Replication Framework yields two important implications for practice in the planning of replication studies. First, although assumptions for the direct replication of results are stringent, it is possible for researchers to address and probe these assumptions through the thoughtful use of research designs and diagnostic tests. A second implication of the framework is that research designs may be used to identify potential sources of effect heterogeneity by systematically violating one or multiple assumptions (while meeting all other assumptions). Here, a prospective design may be applied to address all other replication design assumptions with the exception of the one that is under investigation. If results are found to not replicate, the researcher will know why there was a difference in effects.

The paper highlights real world examples of how research design approaches and empirical diagnostics may be used to conduct high quality replication studies. The first example will come from a series of prospectively planned, highly controlled replication studies in the context of a teacher preparation program; the second is a post-hoc replication of a random assignment field trial offering full- and half-day preschool. Data from these two sets of replication studies will be used to demonstrate various research design approaches for replication, the feasibility of proposed methods on real world applications of replication, and how researchers may address plausible threats to replication design assumptions when they do occur in field settings.

Conclusions and implications (expected):
The paper demonstrates that research designs for replication may be used to systematically evaluate the replicability of effects and identify sources of effect heterogeneity. The paper concludes with a discussion of methodological tools that are needed to conduct high quality replication studies in field settings.

References
Bollen, K., Cacioppo, J., Kaplan, R., Krosnick, J. A., & Olds, J. L. (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Report of the Subcommittee on Replicability in Science Advisory Committy to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.
Hansen, W. B. (2011). Was Herodotus Correct? Prevention Science, 12(2), 118–120. https://doi.org/10.1007/s11121-011-0218-5