Last summer saw the publication of a monumental piece of work: the reproducibility project (Open Science Collaboration, 2015). In a huge community effort, over 250 researchers directly replicated 100 experiments initially conducted in 2008. Only 39% of the replications were significant at the 5% level. Average effect size estimates were halved. The study design itself—conducting direct replications on a large scale—as well as its outcome are game-changing to the way we view our discipline, but students might wonder: what game were we playing before, and how did we get here?

In this blog post, I provide a selective account of what has been dubbed the “reproducibility crisis”, discussing its potential causes and possible remedies. Concretely, I will argue that adopting Registered Reports, a new publishing format recently also implemented in JEPS (King et al., 2016; see also here), increases scientific rigor, transparency, and thus replicability of research. Wherever possible, I have linked to additional resources and further reading, which should help you contextualize current developments within psychological science and the social and behavioral sciences more general.

In 2005, Ioannidis made an intriguing argument. Because the prior probability of any hypothesis being true is low, researchers continuously running low powered experiments, and as the current publishing system is biased toward significant results, most published research findings are false. Within this context, spectacular fraud cases like Diederik Stapel (see here) and the publication of a curious paper about people “feeling the future” (Bem, 2011) made 2011 a “year of horrors” (Wagenmakers, 2012), and toppled psychology into a “crisis of confidence” (Pashler & Wagenmakers, 2012). As argued below, Stapel and Bem are emblematic of two highly interconnected problems of scientific research in general.

PUBLICATION BIAS

Stapel, who faked results of more than 55 papers, is the reductio ad absurdum of the current “publish or perish” culture[1]. Still, the gold standard to merit publication, certainly in a high impact journal, is \( p < .05 \), which results in publication bias (Sterling, 1959) and file-drawers full of nonsignificant results (Rosenthal, 1979; see Lane et al., 2016, for a brave opening; and #BringOutYerNulls). This leads to a biased view of nature, distorting any conclusion we draw from the published literature. In combination with low-powered studies (Cohen, 1962; Button et al., 2013; Fraley & Vazire; 2014), effect size estimates are seriously inflated and can easily point in the wrong direction (Yarkoni, 2009; Gelman & Carlin, 2014). A curious consequence is what Lehrer has titled “the truth wears off” (Lehrer, 2010). Initially high estimates of effect size attenuate over time, until nothing is left of them. Just recently, Kaplan and Lirvin reported that the proportion of positive effects in large clinical trials shrank from 57% before 2000 to 8% after 2000 (Kaplan & Lirvin, 2015). Even a powerful tool like meta-analysis cannot...
clear the view of a landscape filled with inflated and biased results (van Elk et al., 2015). For example, while meta-analyses concluded that there is a strong effect of ego-depletion of Cohen’s d=.63, recent replications failed to find an effect (Lurquin et al., 2016; Sripada et al., in press)[2].

GARDEN OF FORKING PATHS

In 2011, Daryl Bem reported nine experiments on people being able to “feel to future” in the Journal of Social and Personality Psychology, the flagship journal of its field (Bem, 2011). Eight of them yielded statistical significance, p < .05. We could dismissively say that extraordinary claims require extraordinary evidence, and try to sail away as quickly as possible from this research area, but Bem would be quick to steal our thunder.

A recent meta-analysis of 90 experiments on precognition yielded overwhelming evidence in favor of an effect (Bem et al., 2015). Alan Turing, discussing research on psi related phenomena, famously stated that

“These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately, the statistical evidence, at least of telepathy, is overwhelming.” (Turing, 1950, p. 453; cf. Wagenmakers et al., 2015)

How is this possible? It's simple: Not all evidence is created equal. Research on psi provides us with a mirror of “questionable research practices” (John, Loewenstein, & Prelec, 2012) and researchers' degrees of freedom (Simmons, Nelson, & Simonsohn, 2011), obscuring the evidential value of individual experiments as well as whole research areas[3]. However, it would be foolish to dismiss this as being a unique property of obscure research areas like psi. The problem is much more subtle.

The main issue is that there is a one-to-many mapping from scientific to statistical hypotheses[4]. When doing research, there are many parameters one must set; for example, should observations be excluded? Which control variables should be measured? How to code participants’ responses? What dependent variables should be analyzed? By varying only a small number of these, Simmons et al. (2011) found that the nominal false positive rate of 5% skyrocketed to over 60%. They conclude that the “increased flexibility allows researchers to present anything as significant.” These issues are elevated by providing insufficient methodological detail in research articles, by a low percentage of researchers sharing their data (Wicherts et al., 2006; Wicherts, Bakker, & Molenaar, 2011), and in fields that require complicated preprocessing steps like neuroimaging (Carp, 2012; Cohen, 2016; Luck and Gaspelin, in press).

An important amendment is that researchers need not be aware of this flexibility; a p value might be misleading even when there is no "p-hacking", and the hypothesis was posited ahead of time (i.e. was not changed after the fact-HARKing; Kerr, 1992). When decisions are contingent on the data are made in an environment in which different data would lead to different decisions, even when these decisions "just make sense," there is a hidden multiple comparison problem lurking (Gelman & Loken, 2014). Usually, when conducting N statistical tests, we control for the number of tests in order to keep the false positive rate at, say, 5%. However, in the aforementioned setting, it is not clear what N should be exactly. Thus, results of statistical tests lose their meaning and carry little evidential value in such exploratory settings; they only do so in confirmatory settings (de Groot, 1954/2014; Wagenmakers et al., 2012). This distinction is at the heart of the problem, and gets obscured because many results in the literature are reported as confirmatory, when in fact they may very well be exploratory-most frequently, because of the way scientific reporting is currently done, there is no way for us to tell the difference.

To get a feeling for the many choices possible in statistical analysis, consider a recent paper in which data analysis was crowdsourced from 29 teams (Silberzahn et al., submitted). The question posited to them was whether dark-skinned soccer players are red-carded more frequently. The estimated effect size across teams ranged from .83 to 2.93 (odds ratios). Nineteen different analysis strategies were used in total, with 21 unique combinations of covariates; 69% found a significant relationship, while
31% did not.

A reanalysis of Berkowitz et al. (2016) by Michael Frank (2016; blog here) is another, more subtle example. Berkowitz and colleagues report a randomized controlled trial, claiming that solving short numerical problems increase children's math achievement across the school year. The intervention was well designed and well conducted, but still, Frank found that, as he put it, "the results differ by analytic strategy, suggesting the importance of preregistration."

Frequently, the issue is with measurement. Malte Elson—whose twitter is highly germane to our topic—has created a daunting website that lists how researchers use the Competitive Reaction Time Task (CRTT), one of the most commonly used tools to measure aggressive behavior. It states that there are 120 publications using the CRTT, which in total analyze the data in 147 different ways!

This increased awareness of researchers' degrees of freedom and the garden of forking paths is mostly a product of this century, although some authors have expressed this much earlier (e.g., de Groot, 1954/2014; Meehl, 1985; see also Gelman's comments here). The next point considers an issue much older (e.g., Berkson, 1938), but which nonetheless bears repeating.

STATISTICAL INFERENCE

In psychology and much of the social and behavioral sciences in general, researchers overly rely on null hypothesis significance testing and $p$ values to draw inferences from data. However, the statistical community has long known that $p$ values overestimate the evidence against $H_0$ (Berger & Delampady, 1987; Wagenmakers, 2007; Nuzzo, 2014). Just recently, the American Statistical Association released a statement drawing attention to this fact (Wasserstein & Lazar, 2016); that is, in addition to it being easy to obtain $p < .05$ (Simmons, Nelson, & Simonsohn, 2011), it is also quite a weak standard of evidence overall.

The last point is quite pertinent because the statement that 39% of replications in the reproducibility project were "successful" is misleading. A recent Bayesian reanalysis concluded that the original studies themselves found weak evidence in support of an effect (Etz & Vandekerckhove, 2016), reinforcing all points I have made so far.

Notwithstanding the above, $p < .05$ is still the gold standard in psychology, and is so for intricate historical reasons (cf., Gigerenzer, 1993). At JEPS, we certainly do not want to echo calls nor actions to ban $p$ values (Trafimow & Marks, 2015), but we urge students and their instructors to bring more nuance to their use (cf., Gigerenzer, 2004).

Procedures based on classical statistics provide different answers from what most researchers and students expect (Oakes, 1986; Haller & Krauss; 2002; Hoekstra et al., 2014). To be sure, $p$ values have their place in model checking (e.g., Gelman, 2006—are the data consistent with the null hypothesis?), but they are poorly equipped to measure the relative evidence for $H_1$ or $H_0$ brought about by the data; for this, researchers need to use Bayesian inference (Wagenmakers et al., in press). Because university curricula often lag behind current developments, students reading this are encouraged to advance their methodological toolbox by browsing through Etz et al. (submitted) and playing with JASP.

Teaching the exciting history of statistics (cf. Gigerenzer et al., 1989; McGrayne, 2012), or at least contextualizing the developments of currently dominating statistical ideas, is a first step away from their cookbook oriented application.

While we can only point to the latter, statistical issue, we can actually eradicate the issue of publication bias and the garden of forking paths by introducing a new publishing format called Registered Reports. This format was initially introduced to the journal Cortex by Chris Chambers (Chambers, 2013), and it is now offered by more than two dozen journals in the fields of psychology, neuroscience, psychiatry, and medicine (link). Recently, we have also introduced this publishing format at JEPS (see
Specifically, researchers submit a document including the introduction, theoretical motivation, experimental design, data preprocessing steps (e.g., outlier removal criteria), and the planned statistical analyses prior to data collection. Peer review only focuses on the merit of the proposed study and the adequacy of the statistical analyses[5]. If there is sufficient merit to the planned study, the authors are guaranteed in-principle acceptance (Nosek & Lakens, 2014). Upon receiving this acceptance, researchers subsequently carry out the experiment, and submit the final manuscript. Deviations from the first submissions must be discussed, and additional statistical analyses are labeled exploratory.

In sum, by publishing regardless of the outcome of the statistical analysis, registered reports eliminate publication bias; by specifying the hypotheses and analysis plan beforehand, they make apparent the distinction between exploratory and confirmatory studies (de Groot 1954/2014), avoid the garden of forking paths (Gelman & Loken, 2014), and guard against post-hoc theorizing (Kerr, 1998).

Even though registered reports are commonly associated with high power (80-95%), this is unfeasible for student research. However, note that a single study cannot be decisive in any case. Reporting sound, hypothesis-driven, not-cherry-picked research can be important fuel for future meta-analysis (for an example, see Scheibehenne, Jamil, & Wagenmakers, in press).

To avoid possible confusion, note that preregistration is different from Registered Reports. The former is the act of specifying the methodology before data collection, while the latter is a publishing format. You can preregister your study on several platforms such as the Open Science Framework or AsPredicted. Registered reports include preregistration but go further and have the additional benefits such as peer review prior to data collection and in-principle acceptance.

In sum, there are several issues impeding progress in psychological science, most pressingly the failure to distinguish between exploratory and confirmatory research, and publication bias. A new publishing format, Registered Reports, provides a powerful means to address them both, and, to borrow a phrase from Daniel Lakens, enable us to “sail away from the seas of chaos into a corridor of stability” (Lakens & Evers, 2014).

SUGGESTED READINGS

- Twitter (or you might miss out)

FOOTNOTES


[2] Baumeister (2016) is a perfect example of [how not to respond](#) to such a result. Michael Inzlicht shows how to respond adequately [here](#).

[3] For a discussion of these issues with respect to the precognition meta-analysis, see Lakens (2015) and Gelman (2014).

[4] Another related, crucial point is the lack of theory in psychology. However, as this depends on whether you read the Journal of Mathematical Psychology or, say, Psychological Science, it is not
addressed further. For more on this point, see for example Meehl (1978), Gigerenzer (1998), and a class by Paul Meehl which has been kindly converted to mp3 by Uri Simonsohn.

[5] However, it would be premature to put too much blame on \( p \). More pressingly, the misunderstandings and misuse of this little fellow point towards a catastrophic failure in undergraduate teaching of statistics and methods classes (for the latter, see Richard Morey’s recent blog post). Statistics classes in psychology are often boringly cookbook oriented, and so students just learn the cookbook. If you are an instructor, I urge you to have a look at “Statistical Rethinking” by Richard McElreath. In general, however, statistics is hard, and there are many issues transcending the frequentist versus Bayesian debate (for examples, see Judd, Westfall, and Kenny, 2012; Westfall & Yarkoni, 2016).

[6] Note that JEPS already publishes research regardless of whether \( p < .05 \). However, this does not discourage us from drawing attention to this benefit of Registered Reports, especially because most other journals have a different policy.

This post was edited by Altan Orhon.