Good science, bad science: Questioning research practices in psychological research

Bakker, M.

Chapter 8

Discussion: Towards improving psychological science

In this dissertation we have questioned the current research practices in psychological science. We specially focused on the problems with the reporting of statistical results and showed that reporting errors are rather common in the psychological literature (Chapter 2) and related to other questionable research practices (QRPs) like not sharing the data with other researchers for verification purposes (Chapter 3). However, results did not highlight a relationship between reporting errors and the removal of outliers (Chapter 4). At the same time, results of Chapter 4 did show discrepancies between sample size descriptions and statistical results, suggesting that exclusions of data are commonly not reported (see also LeBel et al., 2013). Moreover, we investigated the consequences of applying several commonly used QRPs on Type I error rates and effect size estimates. The use of QRPs like the ad hoc exclusion of outliers to obtain a significant result will lead to substantially inflated Type I error rates (Chapter 5), which increases the probability of publishing false positive results. The use of QRPs also results in biased effect size estimates and distorted meta-analytical results (Chapter 6). These issues are particularly problematic in light of the scarcity of direct replication attempts (Makel et al., 2012) and non-significant results (Sterling et al., 1995) in the psychological literature. Indeed, indications of such biases appeared in half of the meta-analyses analyzed in Chapter 6 (see also: Ferguson & Brannick, 2012). Such signs of bias may cast doubt on some prominent findings in the literature that are highlighted in many psychological textbooks (Appendix C).

Furthermore, we investigated the power paradox, or the question of why the psychological literature contains so many significant results based on underpowered studies. In Chapter 6, we showed that the running of multiple underpowered studies with a small sample size combined with the use of QRPs represents the "optimal" strategy for a researcher if his or her goal is to find a significant *p* value in the hypothesized direction. However, this strategy also resulted in an inflated Type I error rate and biased effect size estimates. Another reason for the power paradox might be the flawed intuitions about power of many researchers. Specifically, results of Chapter 7 showed that researchers strongly overestimated the power of typical studies in their work and that such a bias is even stronger when researchers review the work of their peers.

Taken together, the current results and those of others (e.g., John et al., 2012; Masicampo & Lalande, 2012; Simmons et al., 2011) do not convey a particularly positive image of the psychological literature and the psychological researcher, although it should be noted that similar problems have been highlighted in the medical sciences (Ioannidis, 2005), neurosciences (Button et al., 2013), and many other fields (Fanelli, 2009, 2010; Stroebe, Postmes, & Spears, 2012). Given that the current publication, grant, and tenure systems in (psychological) science strongly stress novel, "exiting", and significant results, the problem has many structural facets that need improvement. In this final chapter, we discuss the current directions and initiatives that are already improving or will hopefully improve research practices in psychological science in the future.

Confirmatory and exploratory research

Most experiments in the psychological literature are presented in the hypothetico-deductive scheme as confirmatory research in which a researcher first derives a testable hypothesis on the basis of theory and subsequently tests this hypothesis empirically. However, it has been argued that this scheme is not followed in a substantial number of publications in which at least some of the hypothesizing takes place after the results are known (Kerr, 1998). For instance, a researcher may setup and conduct a study that includes several loosely defined dependent variables. The researcher subsequently selects and reports only the dependent variable that shows a significant result and derives a hypothesis for that outcome retrospectively. Although such studies are presented as confirmatory studies, they are in fact partly exploratory, since they are not based only on an a priori hypothesis. Exploratory studies are important in their own right (i.e., for finding previous unknown effects and relations). However, both confirmatory and exploratory research should be presented as such, as confirmatory conclusions based on exploratory analyses will lead to unreliable conclusions (Kerr, 1998; Wagenmakers et al., 2012). Therefore, researchers should make a clear distinction between exploratory and confirmatory research (De Groot, 1956; Jaeger & Halliday, 1998; Wagenmakers et al., 2012). To prevent possible strategic behavior and the influence of different biases during the data collection and analysis stage, researchers should preferably preregister their confirmatory study (Wagenmakers et al., 2012). That is, they should register the hypotheses, study design, and data-analysis plan at for example http://openscienceframework.org *before* starting the data collection, and following these plans while analyzing the data. If following of the preregistered plan is not possible (e.g., due to unexpected events) or interesting but not preregistered results show up, researchers could still mention these results, but should clearly label these results as exploratory (see Chapter 4 for an example).

Recently, more than 80 scientists signed an open letter, which is published in *The Guardian*, in which they encourage scientific journals to accept preregistered studies independent of what the results will show (Chambers & Munafo, 2013). More and more journals are starting to answer to this appeal and start experimenting with publishing preregistered studies. For example, the journal *Cortex* is now open for *Registered Reports*, of which the submission will be reviewed prior to data collection, and if accepted, the study will be published independent of what the results show (Chambers, 2013). Other journals that start with preregistration initiatives are *Perspectives on Psychological Science*, *Attention, Perception & Psychophysics*, and *Psychological Science*. Besides making truly confirmatory studies possible and recognizable, these initiatives will also lead to more well powered and better-designed experiments. More importantly perhaps, they will force researchers (and editors and reviewers) more strongly than currently is the case to let the (empirical) chips fall where they may.

Replication

Direct and sufficiently powerful replications are needed to ‘confirm’ results and correct false positives (Asendorpf et al., 2013; Bakker et al., 2013). Until recently only a few direct replications were published in psychology (Makel et al., 2012), since journals typically prefer novel and significant findings. This is problematic because (the publishing of) non-significant results and (direct) replication attempts are important for acquiring a complete picture of empirical results in meta-analyses and systematic reviews. Non-significant results in particular have long been known to be hard to publish (Sterling, 1959) and this situation has not improved in the last decades (Fanelli, 2012). Fortunately, we currently see some positive changes also in this aspect. For example, replication attempts can be posted on http://www.psychfiledrawer.org/, and the journals *Social Psychology* and *Frontiers in Cognition* are currently working on special issues in which only the results of (preregistered) replication studies will be published. Furthermore, the reproducibility project, led by Brian Nosek, investigates the current reproducibility of psychological science by replicating studies published in the 2008 issues of three important psychological journals (Open Science Collaboration, 2012). This project currently involves more than 150 scientists from around the world. The recently completed many labs project (Klein et al., in press) is a comparable project in which the variation in replicability was examined of thirteen classic psychological effects across 36 samples and settings with over 6000 participants. Ten classic effects replicated consistently, one effect showed weak support for replicability, and two effects related to behavioral priming did not replicate. Although these initiatives are some steps forward, the focus of journals on novel and significant findings needs to change to a focus on the quality of the study. *PLoS ONE* tries to achieve this by clearly stating in their information for reviewers: “Unlike many journals which attempt to use the peer review process to determine whether or not an article reaches the level of 'importance' required by a given journal, *PLoS ONE* uses peer review to determine whether a paper is technically sound and worthy of inclusion in the published scientific record.” (PLoS ONE, 2014).

Statistical methods

Although null-hypothesis significance testing (NHST) is still the most common method of statistical inference in psychology (Hubbard & Ryan, 2000; Nickerson, 2000; Sterling et al., 1995), it is not without critics (Cohen, 1994; Cumming, 2013; Wagenmakers, 2007). One of the problems is that the results are often wrongly interpreted, especially the *p* values (Cumming, 2013; Gigerenzer, 2004; Wagenmakers, 2007). Cumming (2012, 2013) therefore proposed to focus on estimation instead of hypothesis testing, and recommends the reporting of effect sizes and confidence intervals. Unfortunately, the interpretation of confidence intervals is also not without problems (Hoekstra, Morey, Rouder, & Wagenmakers, in press).

Another solution is to use Bayesian statistics (Berger, 1985; Wagenmakers, 2007). One advantage of the latter solution is that the evidence for two different hypotheses (e.g., the null hypothesis and the alternative hypothesis) can be quantified in the so-called Bayes factor. The Bayes factor shows under which hypothesis the observed data are most likely and makes it possible to compare both hypotheses directly (Wagenmakers, 2007). Furthermore, while NHST requires a pre-specified stopping rule, optional stopping (the QRP of stopping after one of the two hypotheses has reached a certain predefined evidential threshold) is no problem when using Bayes factors (Dienes, 2011; Wagenmakers, 2007). However, Simmons et al. (2011) argued against the use of Bayesian statistics, since Bayesian statistics will require making additional judgments (e.g., the use of priors), which, according to Simmons et al., creates more opportunities to use questionable research practices (see also Efron, 1986). On the other hand, Wagenmakers (2007) states that inferential procedures that do not take prior knowledge into account are incoherent and may waste possible useful information (see the second online appendix of Wagenmakers, 2007 for a more extensive discussion).

Another problem of NHST as commonly used, is that assumptions of standard parametric tests (e.g., normality, equal variances) are often not met (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000; Micceri, 1989; Wilcox, 2012). These assumptions are often not even checked (Hoekstra et al., 2012), or might be difficult to check (Wilcox, 2012). Other statistical methods like non-parametric and robust methods have less assumptions and are therefore a better solution than parametric test if assumptions are (expected to be) violated (Gibbons & Chakraborti, 2003; Huber, 1981; Wilcox, 2012).

Currently, most introductory statistics textbooks only focus on NHST while not giving much attention to its problems and alternatives. Although this may represent somewhat of a challenge in freshman (undergraduate) courses, at least some attention should be given in (advanced) psychology statistics courses to estimation, robust statistics, and Bayesian inference. For instance, Cumming (2012) introduces estimation, Wilcox (2012) gives an extensive introduction to robust statistics, and Bolstad (2007) provides an introduction to Bayesian statistics. Journals should also be more open to submissions that use statistical methods other than NHST. As good examples, *Psychological Science* and the journals of the Psychonomic Society try to aid researchers in shifting from NHST to estimation, robust statistics or Bayesian inference (Cumming, 2013; Eich, 2013; Psychonomic Society, 2012).

Transparency

Openness of the scientific system is what makes it a successful epistemic project and transparency is therefore required (Wicherts et al., 2012). To improve transparency in psychological science, data sharing should become the standard. Sharing research is not common as only 27% of researchers shared their data (Wicherts et al., 2006). The sharing of data will make the reuse of data possible (e.g., to answer new questions), will force researchers to work more carefully, and will render reporting errors easier to correct (Wicherts & Bakker, 2012). The sharing of data will become more common since online sharing is getting easier at for example the http://opensciencframework.org and data repositories. Also, data sharing is expected to improve because of the requirements by funding organizations like the National Institutes of Health (NIH) that stipulated that all funded research with a certain budget should release and share the data no later than the acceptance for publication (NIH, 2013). Luckily, it appears that more and more (mostly young) researchers are beginning to share their data via online repositories.

Furthermore, reporting standards of psychological research should be improved by implementing clear editorial guidelines and by checking thereof by editors and reviewers. Transparent reporting enables a better evaluation of the research findings. Some additional recommendations based on the research described in this dissertation are: (1) To prevent reporting errors (Chapter 2-4) co-authors should always check the correct reporting of the results by running the analyses independently. (2) The power of psychological studies is often low. This might be due to strategic consideration (Chapter 6), but also because of flawed intuitions about power (Chapter 7). Even researchers who state that they typically conduct a power analysis to determine sample size had poor intuitions about power. Therefore, authors should be more transparent about their sample size decisions, by reporting a formal and a priori power analysis, and authors should explain the use of sample sizes that have low power to detect the effect of interest. (3) Data exclusions should be always reported, and preferably an outlier-handling plan is part of the preregistration (see Chapters 4 and 5). (4) As stated above, results from confirmatory and exploratory analyses should be clearly distinguished in the manuscript. Preregistration of confirmatory studies and analyses is the best antidote against QRPs, publication bias, and the all too human tendency of researchers to adapt analytic plans (or alter the hypotheses) depending on the results.

To improve transparency, LeBel et al. (2013) set up PsychDisclosure.org on which authors can disclose design specifications for four methodological categories (exclusion, non-reported conditions and measures, and sample size determination). Simmons et al. (2012) proposed to make reports more transparent by including (if applicable) the following statement to a manuscript “we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in this study.” Both initiatives will improve transparency and have inspired some of the changes in the publication standards and practices in *Psychological Science* that were introduced in 2014 (Eich, 2013). Also the Psychonomic Society (2012) has new guidelines for authors that will improve the transparency of the articles in their journals. Hopefully other journals will follow shortly.

The future

Psychological science is a complex system that is often highly competitive and involves many different actors, including researchers, institutions, professional organizations, journals, publishers, and funding organizations. The current crisis in psychology has highlighted some real weak spots in our scientific enterprise, which may have led to a distorted picture of results in the scientific literature. Fortunately the current discussion (e.g., special issue on replicability of *Perspectives in Psychological Science* of which Chapter 6 was part) has also led to major strides forward. Some changes are top down and originate from funding organizations (e.g., the obligation to share NIH funded research data) and journal editors (e.g., the new journal policies of *Psychological Science* and the Psychonomic Society, and the initiatives of different journals to experiment with publishing preregistered studies). Other initiatives are bottom up, like the call of many scientists for more preregistered studies, and the different collaborative replication projects. These initiatives drew many scientists from different research fields and different parts of the world together, and will hopefully lead to improved research practices and a flourishing psychological science.