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

Bakker, M.

Chapter 1

Introduction

Psychology in crisis

In 2011 several events shook up psychology, raising questions about the credibility of psychological research (Asendorpf et al., 2013; Carey, 2011; Pashler & Wagenmakers, 2012). A much-debated event was the extensive fraud by social psychologist Diederik Stapel, who admitted to having fabricated data over a prolonged period in his prolific career. A thorough investigation by the Levelt committee (Levelt Committee, 2012) established fraud in 55 publications by Stapel, and these publications are (or will be) retracted. Another and perhaps more thought-provoking event was the publication of a controversial article on precognition by Daryl Bem (2011) in a flagship journal in social psychology. In this article in the *Journal of Personality and Social Psychology (JPSP)*, Bem reported the results of nine experiments that aimed to show effects of manipulations despite the fact that the outcomes were measured *before* the manipulations had taken place. The editors of JPSP stated that they accepted Bem's paper on the basis of the regular standards in place at the journal (Judd & Gawronski, 2011), at which 85% of the submissions were rejected in 2011 (American Psychological Association, 2012). In this article, Bem (2011) supposedly 'proved' the existence of precognition by presenting significant effects in eight of the nine experiments. These findings are controversial even to the editors, who stated that "they turn our traditional understanding of causality on its head" (Judd & Gawronski, 2011, p. 406).

Null hypothesis significance testing

According to Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) Bem's paper illustrates deep-rooted problems with the way psychological researchers analyze their data and present their results. Although it remains controversial (Cohen, 1994; Cumming, 2013; Wagenmakers, 2007), the most widely used inferential method used by psychological researchers is null hypothesis significance testing (NHST; Cohen, 1962, 1994; Hubbard & Ryan, 2000; Maxwell, 2004; Nickerson, 2000; Sterling, Rosenbaum, & Weinkam, 1995). The null hypothesis represents the hypothesis of no difference between different sets of data or of no effect of an experimental manipulation on the dependent variable(s) of interest. NHST provides *p* values, which, if smaller than a predefined threshold (typically $\alpha = .05$), lead to the rejection of the null hypothesis of no difference or no effect (Nickerson, 2000). Such significant outcomes, in turn, are typically seen as lending support for the alternative hypothesis, or the hypothesized (non-null) difference or effect (Hoekstra, Finch, Kiers, & Johnson, 2006; Mahoney, 1977; Rosenthal & Gaito, 1963).

An α of .05 means that the researcher accepts a chance of 5% of wrongfully rejecting the null hypothesis, leading to the conclusion that effect exist when in fact it does not. This is also called a false positive or a Type I error. Many scientists and lay people would probably consider precognition nonexistent and so would consider the eight significant effects reported by Bem (2011) Type I errors. Bem could have been extremely lucky to find a false positive 8 out of 9 times, or something else might be going on. One explanation is that many different studies have been conducted and only the (expected) 5% Type I errors were included in Bem's manuscript. According to this view, the (expected) 95% studies in which the null hypothesis could not be rejected ended up in the proverbial file drawer (Rosenthal, 1979).

Questionable research practices

Another explanation of the occurrence of potential Type I errors might be the use of Questionable Research Practices (QRPs), which are research practices that fall in the grey area, like continuing data collection until a statistically significant effect is found (or sequential testing; e.g., Wagenmakers, 2007). In 2011 too, Simmons, Nelson, and Simonsohn showed that the use of QRPs in the collection and analysis of data can lead to dramatically inflated Type I error rates. They eloquently illustrated this by presenting a study that 'proved' (on the basis of a significant outcome) that listening to "When I'm sixty-four" by The Beatles makes people nearly a year-and-a-half younger. The study was genuine, but by using several QRPs (or researcher degrees of freedom as they called it) Simmons et al. were able to test different combinations until a significant effect showed up. Subsequently, Simmons et al. showed in a simulation study that a combination of QRPs (using multiple potentially useful dependent variables, adding 10 observations after an intermediate test failed to provide a significant outcome, controlling for covariates, and ad hoc dropping of conditions) led to a Type I error rate of 61% instead of the generally accepted rate of 5%. Simmons et al.'s results showed that Type I error rates can be inflated strongly by the use of QRPs, but did not show how commonly used these practices were.

John, Loewenstein, and Prelec (2012) investigated the prevalence of the use of different QRPs by surveying anonymously more than 2,000 psychological researchers on whether they had ever used QRPs in their work. Examples of QRPs in their survey were not reporting all of a study's dependent variables (admitted by 63% of the respondents) and deciding whether to collect more data after looking to see whether the (intermediate) results were significant (admitted by 56% of the respondents). Also, nearly half of the respondents indicated that they had selectively reported in their paper only those studies that "worked", which supposedly refers to studies with desirable (significant) results. If indeed QRPs are as widely

used in psychological research as John et al.'s results would suggest, the actual Type I error rate in psychological research surpasses the nominal level of .05. The overarching goal of this dissertation is to study the use of QRPs in the context of NHST in psychology. We do not debate NHST per se, but note that it continues to be the most popular statistical procedure in psychological research and that its use in practice deviates from how it is described in the textbooks (Nickerson, 2000).

p values

Another QRP in John et al.'s (2012) survey concerned the incorrectly 'rounding off' of *p* values ("e.g., reporting that a *p* value of .054 is less than .05", p. 525), which was admitted by 22% of the respondents. Incorrectly rounding off can be detected by comparing the reported *p* value with the recalculated *p* value based on the reported test statistic and degrees of freedom (e.g., $F(1,33) = 4.00$). In Chapter 2 we study the discrepancies between reported *p* values and recalculated *p* values (i.e., reporting errors) in the psychological literature. Results show that approximately half of a large sample of articles in the psychological literature contained a reporting error. In addition, we found that in 15% of the articles a reporting error changed the statistical conclusion. Almost all the errors that changed the conclusion resulted in a recalculated *p* value greater than .05, while the result was reported with a *p* value less than .05. The high prevalence of reporting errors has already been replicated by others (Caperos & Pardo, 2013; Leggett, Thomas, Loetscher, & Nicholls, 2013). Furthermore, Masicampo and Lalande (2012) showed a peculiar prevalence of *p* values just below .05, which can be seen as evidence of so called *p* hacking, or the use of QRPs in the collection and analysis of data to achieve significant results (Simonsohn, Nelson, & Simmons, in press).

In Chapter 3 we relate the good research practice of data sharing with these reporting errors and strength of evidence (against the null hypothesis of no effect) in a set of 48 psychological articles. Results show that articles from which data were shared contained fewer reporting errors and also fewer reporting errors that changed the statistical conclusion than articles from which data were not shared. Furthermore, the median *p* value was lower in articles from which data were shared compared to articles from which data were not shared.

Outliers

Now and then, almost all researchers are confronted with the question what to do with outliers in the data. One of the QRPs from the survey of John et al. (2012) concerned the exclusion of data from the analysis after looking at the impact of doing so on the results. Around 38% of the respondents admitted to having used such practices. The ad hoc exclusion of data is

a clear example of a QRP as both removal and non-removal of an extreme value can have a profound effect on the outcome of an analysis. In Chapter 4, we relate the removal of outliers to reporting errors and strength of evidence (against the null hypothesis of no effect). We followed the same procedure as in Chapter 3 and were therefore able to preregister our hypotheses and methods. However, we did not find any differences between articles in which outliers were removed and articles in which outliers were not removed, but found evidence of contamination of our sample due to non-transparent reporting of the exclusion of data.

In Chapter 5 we investigate more closely the impact of outlier removal on the Type I error rate in *t* tests. After investigating the common practice of outlier handling in the psychological literature, we show by means of a simulation study the inflation of the Type I error rate of independent samples *t* tests, when outliers are removed based on commonly used *Z* value thresholds from realistic datasets based on sum scores. Furthermore, we show that non-parametric and robust tests are a good replacement of the independent samples *t* test, as these methods show a nominal Type I error rate and comparable or better power (the probability of correctly rejecting the null hypothesis) when outliers are either absent or present in the data.

Replications

In their critique of Bem's (2011) article on precognition, Wagenmakers et al. (2011) noted that Bem (2000) himself recommended in a textbook on scientific publishing that researchers "go on a fishing expedition for something-anything-interesting" (p. 5). This so called fishing expedition includes the use of QRPs, and Wagenmakers et al. argued that Bem had used these QRPs in his work on precognition, while presenting his results as confirmatory. The controversy related to publication of Bem's (2011) study in a flagship journal was not limited to how Bem analyzed his data. Several later replication attempts failed to find significant results in support of Bem's claims (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012). Publishing these replications was not easy. For example, editors at *JPSP*, *Science*, and *Psychological Science* all stated that these journals did not publish direct replications of earlier work and did not even send the paper of Ritchie et al. out for review. The *British Journal of Psychology* sent the paper out for review, but rejected the paper, since Bem was one of the reviewers (see Wicherts, Kievit, Bakker, & Borsboom, 2012 for more discussion of problems with peer-reviews of critical articles). In the end, Ritchie et al. published their paper in *PloS ONE* (Yong, 2012). Direct replication studies, in which the original protocol is followed as closely as possible, are important to correct published false positives (i.e., Type I errors; Asendorpf et al., 2013; Schmidt, 2009). Nevertheless, Makel, Plucker, and Hegarty (2012)

showed that only 1.07% of a large sample of articles in psychology contained a replication study of which 18% involved direct replications.

The most common argument by reviewers and editors to not consider publishing replications is that successful replications are not that interesting, as they do not present something new (Neuliep & Crandall, 1990, 1993). Unsuccessful replications are difficult to publish because non-significant results (i.e., the null hypothesis is not rejected) are in general difficult to publish (Fanelli, 2010; Greenwald, 1975; Sterling, 1959; Sterling et al., 1995). One of the reasons why it is difficult to publish non-significant results is that these studies also do not support the null hypothesis since other explanations might apply (e.g., the power to detect an effect might be too low or the manipulation might have gone wrong). In a survey, Greenwald (1975) found that both researchers and reviewers prefer results in which the null hypothesis could be rejected. When the null hypothesis was rejected, 49% of the respondents would submit the results for publication before further data collection, whereas only 6% would do this when the null hypothesis was not rejected. Furthermore, Mahoney (1977) showed with an experiment that reviewers who reviewed a manuscript in which the results were not statistically significant rated the methodology and data presentation lower than those who reviewed the same manuscript but with statistically significant results.

This preference for publishing only new and significant results will lead to a bias in the published literature, also called publication bias (Rothstein, Sutton, & Borenstein, 2005). According to Ioannidis (2005) this leads to a situation in which most published research findings in the medical sciences are false positives. Publication bias is especially problematic for meta-analyses in which the results of several studies are aggregated, because the published studies might be false positives or might have inflated effect sizes if the population effects deviate from zero (Ioannidis, 2008b). Therefore, the aggregated results in a meta-analysis might give a distorted picture of an effect. In Chapter 6 we show with a reanalysis of 13 meta-analyses the problems of publication bias and with a simulation study the effect of publication bias combined with the use of QRPs on the results of meta-analyses.

Human factors

More than 90% of the papers involving NHST have been shown to report a significant effect (Fanelli, 2010; Sterling et al., 1995). This is peculiar because psychological effects are typically not large (Anderson, Lindsay, & Bushman, 1999; Hall, 1998; Lipsey & Wilson, 1993; Meyer et al., 2001; Richard, Bond, & Stokes-Zoota, 2003; Tett, Meyer, & Roese, 1994) and most psychological publications contain (multiple) small and therefore underpowered studies (Cohen, 1990; Francis, 2012b; Maxwell, 2004; Schimmack, 2012). One explanation of this

power paradox is that conducting multiple underpowered studies is the most efficient strategy for a researcher, in terms of finding a significant and therefore publishable effect. We show this with a simulation study in Chapter 6 in which we also show that this strategy leads to inflated Type I error rates and biased effect size estimates.

This explanation of the strategic researcher, pictures him or her as a cunning scientist who is mostly interested in publishing, independent of the reliability of the content of his or her publications. More likely, the researcher is just as all other people sensitive to different biases and other human factors (Mahoney, 1979). QRPs might be applied without awareness of the consequences for the Type I error rate of the specific study and science in general. To prevent experimenter expectancy effects (Rosenthal, 1966), experimenters are preferably unaware of the study's aim and are 'blind' to the condition the participants are (randomly) assigned to. The same does not apply to the researcher who analyses the data. The statistical analyses are often conducted by a person who is (1) aware of the study's main hypothesis, (2) who is likely to believe this hypothesis, and (3) who often benefits directly from finding support for it. This is not an ideal mix for objective and sound results as the decisions made by researchers (e.g., handling of outliers, choice of analysis) may lean towards the goal of achieving a (significant) result in line with the proposed hypothesis (Barber, 1976; Friedlander, 1964).

There is some evidence that scientists' reaction to data depends on whether or not these data support their hypotheses (Edwards & Smith, 1996; Fugelsang, Stein, Green, & Dunbar, 2004; Koehler, 1993; Mynatt, Doherty, & Tweney, 1977). For instance, Fugelsang et al. (2004) interviewed molecular biologists concerning their reactions to empirical results that were at odds with their hypothesis. Unexpected results were generally dismissed on methodological grounds, while results that matched their hypothesis were easily accepted. Similarly, Edwards and Smith (1996) argued that people have a tendency to reject data that contradicts their hypotheses and search harder for flaws in such data. Searching for evidence that confirms your preexistent beliefs is called confirmation bias (Nickerson, 1998) and rejecting evidence that contradict preexistent beliefs is called disconfirmation bias (Edwards & Smith, 1996). This means that researchers hold double standards with respect to the quality of their data and analyses. If the data and the (planned) first analyses support the favored hypothesis, then researchers will show little scrutiny with regards to the data and the analyses. This might result in errors being overlooked. On the other hand, if the data and the (planned) analyses fail to support the cherished hypothesis, researchers are likely to check for errors. For example, this might be a reason why almost all of the reporting errors that changed the statistical conclusion resulted in a significant p value (Chapter 2). Rosenthal (1978) also showed that two thirds of recording errors favored the hypothesis of the researcher. Another way these biases might work is that

studies that reject the null hypothesis are accepted without ample consideration, while studies that do not reject the hypothesis are extensively evaluated until some problems in the design come forward and the study is declared to be 'failed'. Schimmack (2012) states this as follows: "In this way, empirical studies no longer test theoretical hypotheses because they can only produce two results: Either they support the theory ($p<.05$) or the manipulation did not work ($p>.05$)" (p. 554).

Another type of bias is the statistical bias. In their classic study Tversky and Kahneman (1971) showed that even quantitatively oriented psychologists regard a sample randomly drawn from a population as highly representative of the populations in all essential characteristics, and call this intuition the believe in the law of small numbers. For example, psychologists underestimated the number of participants that is needed to replicate a surprising result. And when the replication failed, most respondents stated that the researcher should try to find an explanation for the difference between the two samples, although the result is quite reasonable when the surprising result is true and power of the studies is low (i.e., the original result and the result of the replication do not differ significantly from each other).

To see whether statistical bias and intuitions might explain the power paradox partly, we present in Chapter 7 the result of our investigation to the intuitions of psychological researchers about power. We surveyed 291 psychological researchers and found that the preferred amount of power (.8 as recommended by Cohen, 1992) is twice the power based on their typical sample size, effect size, and level of α. Participants who indicated that they typically base their sample size decision on a formal power analysis did not have better power intuitions. Also sub-field or seniority had no relation with power intuitions. However, participants who answered the questions from a reviewer's perspective showed even worse intuitions and preferred multiple small studies over one large study.

Since the beginning of the crisis in psychology in 2011, several results (including those in the chapters reported here) have highlighted the problems faced by psychology. At the same time, the field witnessed major strides forward that we discuss in the final chapter (Chapter 8).