Journal of Parapsychology
2021, Vol. 85, No. 1, 108-114

http://doi.org/10.30891/jopar.2021.01.09

The Mote in Thy Brother's Eye

Chris Roe[1]

The University of Northampton

A Review of *The 7 Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*, by Chris Chambers. Princeton, NJ: Princeton University Press, 2017. Pp. 274. $19.95. ISBN: 978-691-15890-7

Chris Chambers has produced a book that provides a welcome response to some of the problems inherent in data collection, analysis, and reporting that have plagued modern psychology. While surely this is a worthy cause, the author adopts a quasi-religious framework (in terms of sins and redemption) that some may find jarring and misjudged, and he does have a tendency to overegg his pudding with generous doses of melodrama. He describes this book as being "borne out of … a deep personal frustration with the working culture of psychological science … if we continue as we are then psychology will diminish as a reputable science and could very well disappear" (p. ix), which derives from a sense that "like so many other 'soft' sciences, we found ourselves trapped within a culture where the appearance of science was seen as an appropriate replacement for the practice of science" (p. ix). Stirring words intended to energise us into action.

He starts badly. In Chapter 1 ('The sin of bias') Chambers surprisingly begins by focusing on Daryl Bem's (2011) publication of a suite of precognition experiments in the *Journal of Personality and Social Psychology* as the point that changed psychology forever by drawing attention to its deeply flawed nature — not the dramatic high profile revelations of fraud perpetrated by psychologists such as Diederik Stapel or Jan Smeesters, nor the dismal failure of the Open Science Replication project, nor the discovery that many psychologists admitted to questionable research practices such as p-hacking and HARKing (Fanelli, 2009), despite all of these being considered in detail later on in the book. No, it was the capacity for a respected psychology journal to entertain the possibility that psi effects could be empirically demonstrated that acted as cue that something had gone very seriously awry. Chambers quickly disparages Bem's data as "nonsensical" and self-evidently untrustworthy. He draws comfort from failed replication attempts by Ritchie, French, and Wiseman, which provide a sober correction of this strange anomaly. The logic here is rather puzzling given that Bem's original 9 experiments involved 1,050 participants, while these replications consisted of just 3 experiments focused on one protocol and tested just 50 participants each. No mention is made of Bem, Tressoldi, Rabeyron, and Duggan's (2016) report on 90 experiments from 33 laboratories in 14 countries that yielded an overall effect greater than 6 sigma ($p = 1.2 \times 10^{-10}$). In a final swipe, Chambers quotes Wagenmakers et al. who saw Bem's publication as an

1 Address correspondence to: Chris Roe, Centre for Psychology & Social Sciences, The University of Northampton, University Drive, Northampton NN1 5PH, email:chris.roe@northampton.ac.uk

indication "that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results" while neglecting to mention two articles by Bayesian statisticians that were critical of their analysis and conclusions (Bem, Utts, & Johnson, 2011; Rouder, & Morey, 2011). Incredibly, all this opens a chapter intended to 'explain how unchecked bias fools us into seeing what we want to see' — I fear the irony may be lost on Chambers, and I confess that after such an inauspicious start I did not have high hopes regarding the merit of this book.

He goes on to explore the 'sin of bias', focusing particularly on confirmatory bias in the refereeing process, which privileges weaker studies that produce unambiguous narratives built around statistically significant results over stronger studies with more nuanced interpretations of a more heterogeneous collection of significant and nonsignificant findings, despite the latter being more likely to reflect the complexities of life outside the laboratory.

Chapter 2 considers 'The sin of hidden flexibility' and introduces the notion of HARKing (Hypothesising After Results are Known) and other questionable research practices (QRPs). He elaborates, 'faced with the career pressure to publish positive findings in the most prestigious and selective journals, it is now standard practice for researchers to analyse complex data in many different ways and report only the most interesting and statistically significant outcomes ... any conclusions drawn from such tests will, at best, overestimate the size of any real effect. At worst they could be entirely false' (p. 23). He links this to p-hacking; that is, the inflation of the alpha level (likelihood of committing a Type I error) to much greater than .05 by the use of multiple independent analyses and selecting post facto only those that suggest effects (see Simmons, Nelson & Simonsohn, 2011), but it extends to many other analytical decisions (such as how to deal with outliers, how to define the DV, whether to include covariates, etc.). He claims,

> in even the simplest experimental design, these [decision] pathways quickly branch out to form a complex decision tree that a researcher can navigate either deliberately or unconsciously in order to generate statistically significant effects. By selecting the most desirable outcomes, it is possible to reject H0 in almost any set of data (p. 25).

A striking example of this compound effect can be seen in Steegen, Tuerlinck, Gelman and Vanpaemel's (2016) multiverse analysis of data from a study of the effects on fertility of religiosity and political attitudes. The dramatic variation in the outcome and conclusions drawn as a result of quite subtle changes to various analytic decisions is sobering.

Recourse to QRPs is claimed to be very common. Chambers refers to remarkable survey data from John, Loewenstein, and Prelec (2012) that suggest a very high proportion of experimental psychologists have on at least one occasion selectively excluded data or failed to report all conditions in an experiment. He concludes that 'far from being a rare practice, p-hacking in psychology may be the norm' (p. 29), but in doing so makes the same error as Bierman, Spottiswoode, and Bijl (2016) in mistaking the proportion of individuals who *ever* engaged in a behaviour for a measure of the behaviour's prevalence (i.e., how frequently they might engage in that practice) — if someone in their childhood once stole on impulse a penny candy, for which they have ever since been ashamed, it does not make them a seasoned shoplifter. Bierman et al.'s QRP model of ganzfeld outcomes achieves optimal fit by assuming prevalences (when practically possible) for QRPs such as relegating confirmatory studies to pilot status

(49%), promoting pilot studies to confirmatory ones (47%), optional stopping (32%), optional extension (44%), publication bias (58%), and post-hoc data exclusion (41%) that in my view are so unrealistic as to be without value; for example, they are orders of magnitude bigger than those suggested by Fiedler and Schwarz (2016) using a more incisive and valid measure (ranging from less than 1% to 10%).

It is not enough to rely on subsequent replication attempts to verify whether novel effects are real or a product of type 1 error (or QRPs), since those studies also can be susceptible to the same decision biases when conducting and reporting analyses. This is especially so when the replication is 'conceptual' rather than exact, whereby some finding is deemed to be broadly consistent with the original finding in ways that allow for shoe fitting. Chapter 2 ends by proposing a set of methods for countering the effects of such 'hidden flexibility', including pre-registration, disclosure statements, data sharing, controlling optional stopping, and community consensus over standard research practices (in terms of appropriate IVs and DVs, rules for dealing with outliers, standard analyses, etc.).

Chapter 3 considers 'the sin of unreliability'. Replication is described as "the immune system of science" (p. 47), acting to identify and neutralise deviant findings, but Chambers recognises that in practice it rarely occurs because the mechanisms of science (notably funding and publication) don't reward direct replications, preferring novelty and innovation (I discuss this in more depth in Roe, 2016b). This section seems to me to retain a naive notion of replication on demand in which any competent person should be able to confirm a published result. Interestingly, the supporters he cites typically come from the natural sciences, where those assumptions might hold. In the social sciences, where participants may be much more sensitive to subtle changes in experimental conditions (including the demeanour of the researcher who interacts with them) there is a much stronger case to be made for experimenter-linked effects (see Roe, 2016c).

A major contributor to replication failures is lack of statistical power. This lesson will be very familiar to those working in parapsychology, thanks to the valuable work of Jessica Utts (e.g., 1999), and don't require retelling here. Chambers bemoans the failure of researchers to adequately describe the methods they have used, which is often attributed to space constraints when publishing papers and a lack of concern among reviewers about 'unnecessary detail'. This can contribute to failures to fully replicate conditions but can also disguise some QRPs such as omitting to mention variables (or even conditions) that didn't work out as expected. Apart from these gross deviations from good practice, however, it seems to me much more complicated in practice to decide what factors need to be described — in how much detail should we describe any preparations we make to be in the right frame of mind as researchers, or to explain our efforts to establish rapport with participants? What might seem to some an unnecessarily meticulous description of prevailing conditions for a supposedly robust psychological effect might seem to others like a genuine effort to capture a complex, subtle human-to-human interaction (see, e.g., Watt, Wiseman & Schlitz, 1998).

In Chapter 4, Chambers moves on to 'The sin of data hoarding', which he characterises as follows:

> Many psychologists consider sharing data only where doing so brings professional gains, such as working partnerships that lead to joint authorship of papers ... They fear that unfettered

> access to data would be to surrender intellectual property to rivals, allow undeserving competitors to benefit from our own hard work and invite unwelcome attention from critics. [In a climate of concern about QRPs] ... who would want their mistakes or dubious decision making exposed to the scrutiny of a rival researcher or (worse) a professional statistician? (p. 76)

Undoubtedly Chalmers is right to draw attention to the many advantages of data sharing, mitigating against error and dubious practices as well as ensuring that data are available to future researchers with previously unthought-of questions or approaches to data analysis. However, there are thorny ethical issues associated with unconstrained data usage such as breaching the terms under which participants originally agreed to provide data, and this requirement might only be achievable prospectively. Additionally, enabling secondary data analysis could actually encourage the kinds of QRP previously discussed in this book. Given access to large amounts of data without the constraints of having to pre-specify hypotheses in the design phase in the way that the study's originators are, and with less time and energy diverted to actually conducting the study and collecting the data, the secondary data analyst is much better placed to subject the data to all sorts of interrogation without feeling obligated to report on all their false trails and missteps as they search for something more interesting to build a journal paper around. One protection against this would be to adopt a policy that all secondary data analysis also needs to be pre-registered.

In practice, data sharing is still very rare. This is starkly illustrated by Wicherts, Borsboom, Kats, and Molenaar's (2006) attempt to secure data from 249 studies published by the American Psychological Association. After sending 400 email requests over 6 months, they had still only received data from 64 studies (25.7% of the total). In support of concerns about dubious practices, Wicherts, Bakker, and Molenaar (2011) found that errors in reporting p values were twice as common in papers by authors who refused to share data than in those where data were provided on request. In a number of cases these errors led nonsignificant outcomes to be reported as significant.

Chapter 5 introduces 'The sin of corruptibility'. Chambers begins with a salutary tale of an early career researcher who travels the slippery slope from playing with data to see what effects analysis decisions might have on the resulting significance value for a putative effect (p-hacking) to outright data manipulation. Although this account is fictitious, he quickly moves on to similar documented cases such as Diederik Stapel (see Roe, 2016a for further details of this case and a comparison with cases from parapsychology). Unfortunately, Chambers succumbs to the temptation to portray Stapel as inadequate, needy and deviant, in a process of othering (in the language of Husserl) that dangerously immunises the wider community — the failings become personal rather than systemic. As the chapter progresses, however, the emphasis shifts to institutional processes. He nicely contrasts, for example, the fairly standard practice of a journal that might review ten papers on the same subject and accept for publication only the two that report statistically significant effects with the dubious practice of a researcher who conducts ten experiments and decides only to submit for publication the two that gave significant outcomes, leaving the others to languish in a file drawer.

Chapter 6 ('The sin of internment') focuses on the restricted way that the results of research are disseminated, criticising the limited access afforded to the general public. Full open access would allow

readers to copy, distribute and display published work in accordance with the particulars of a creative commons attribution licence, often with the result that the financial burden for review, production and publication are carried by the authors rather than readers. This may not be a viable model for many journals, but hybrid approaches are possible, with some 'shop window' articles being free to access while others can be accessed via a paywall. While some journals have opted for these hybrid approaches (such as the APA's journals) psychology as a whole has been slow to shift from restricted access publishing. Chambers attributes this partly to concerns that making pre-publication versions publicly available might draw attention to discrepancies between them and the final published version, which could reflect QRPs such as reframing the aims and hypotheses so they fit better with the reported analyses, but this seems highly speculative. A more likely explanation, also offered here, is that restricted access is in the interest of the highest impact journals and therefore of researchers for whom journal reputation and impact is an important consideration for career advancement. This restriction has severe and unexpected consequences, for example in disadvantaging researchers based in poorer countries, or in making one's work unavailable to policy makers and other key stakeholders (astonishingly, the Higher Education Funding Council for England, which oversees the activity of universities, has 'zero access to non Open Access content' (p. 140)).

Interestingly, articles published in open access journals are cited more frequently than articles with restricted access. Making primary material publicly available would allow the interested neutral to make their own judgements concerning study quality and the plausibility of counter explanations, rather than having to rely on secondary sources. Clearly, a shift toward greater transparency and openness is potentially of great benefit to parapsychology.

The final sin, 'Bean counting', features in Chapter 7. Here, Chambers criticises the metrification of research, which has led to certain parameters (such as grant income and number of outputs) being treated as indicators of research quality and impact. This clearly favours some (expensive) lines of research such as neuroscience over inexpensive lines of research such as qualitative social psychology. Citation indices are seen as particularly pernicious, not least because they are unrepresentative, with 20% of published papers being responsible for 80% of citations (the so-called 80/20 rule). The criteria for calculating impact factors also turns out to be surprisingly susceptible to lobbying and negotiation. Chambers concludes, "whether there is any better metric than JIF [Journal impact factor] is unclear, but it is hard to imagine anything worse" (p. 155).

In the final chapter ('Redemption'), Chambers recaps on the main sources of bias, error, and fraud that have been highlighted throughout the book. He offers solutions that reflect recent initiatives to encourage preregistration and open data. He is particularly keen to promote peer review of papers at the design stage (that is, before data collection has begun), with journal reviewers making recommendations based on rationale and methodology rather than outcomes. Chambers refers to this as a Registered Report, and he notes that while others have made similar suggestions in the past (notably Robert Rosenthal in the 1970s), these have never previously been implemented (p. 179). He is clearly oblivious of the *European Journal of Parapsychology*'s policy regarding pre-acceptance, also dating from the 1970s.

Much of the chapter is devoted to lobbying for the adoption of registered reports. A number of additional proposals, such as producing a 'reproducibility index' seem too labour intensive to gain traction, but others such as more co-ordinated multicentre replication attempts, and much better protection for 'whistle blowers' who identify fraudulent activity by colleagues, seem more promising as a model for good practice in parapsychology.

In summary, *The 7 Deadly Sins of Psychology* is elegantly written, very well researched, and clearly has been produced by someone at the heart of the movement for change. There are fulsome endnotes but unfortunately no separate references list. The recommendations are, on the whole, very sensible; indeed, the case for change seems to me compelling. It's just a pity that Chambers didn't take the time to follow his own advice when it comes to his pronouncements about parapsychology generally and the Bem paradigm in particular. Nevertheless, I would highly recommend the book to anyone conducting research, or having an academic interest, in the social sciences (including parapsychology).

References

Bem, D.J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425.

Bem D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*:1188 doi: 410.12688/f1000research.7177.2

Bem, D.J., Utts, J., & Johnson, W.O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716-9.

Bierman, D.J., Spottiswoode, J.P., & Bijl, A. (2016). Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology. *PLoS ONE 11*(5): e0153049. doi:10.1371/journal.pone.0153049

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE 4*(5), e5738. doi:10.1371/journal.pone.0005738

Fiedler, K., & Schwarz, N. (2015). Questionable research practices revisited. *Social Psychological and Personality Science, 7,* 45-52 doi*: 10.1177/1948550615612150*

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524-532. https://doi.org/10.1177/0956797611430953

Lakens, D. (2014). What p-hacking really looks like: A comment on Masicampo & LaLande (2012). [Blog post]. Retrieved from http://daniellakens.blogspot.com/2014/09/what-p-hacking-really-looks-like.html

Roe, C.A. (2016a). Oh, what a tangled web we weave, when first we practise to deceive: The problem of fraud in parapsychology. *Mindfield, 8*(1), 8-17.

Roe, C.A. (2016b). Is inconsistency our only consistent outcome? *Mindfield, 8*(2), 70-75.

Roe, C.A. (2016c). Experimenter as subject: What can we learn from the experimenter effect? *Mindfield*, *8*(3), 89-97.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689. doi:10.3758/s13423-011-0088-7

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Steegen, S., Tuerlinck, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702-712. https://doi.org/10.1177/1745691616658637

Utts, J. (1999). The significance of statistics in mind-matter research. *Journal of Scientific Exploration*, *13*, 615–638.

Watt, C., Wiseman, R. & Schlitz, M. (1998). Tacit information in remote staring research: The Wiseman-Schlitz interviews. *Paranormal Review, 24*, 18-25. doi: 10.1037/0003-066X.61.7.726

Wicherts, J.M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728.

Wicherts, J.M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE 6*(11): e26828. doi:10.1371/journal.pone.0026828