Behaviour Research and Therapy 116 (2019) 90–93

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Refining our research practices in clinical science: Challenges and steps towards solutions

Questionable research practices and problems with reproducibility in psychology and related fields have been discussed and written about widely (e.g., John, Loewenstein, & Prelec, 2012; Nosek et al., 2015; Open Science Collaboration, 2015). From the 2015 Open Science Collaboration report of psychological science, only 36% of findings reported in top tier journals, mostly pertaining to social and cognitive psychology, were statistically significant in independent replication studies (Nosek et al., 2015). This is not to say that replication is only a psychology issue; no field is immune, from economics (Camerer et al., 2016) to clinical medicine (Ioannidis, 2016). Nor is it an issue of only specialized journals, since 38% of social-science papers published in high-profile journals such as *Nature* and *Science* fail to replicate (Camerer et al., 2018). In a survey by *Nature* (Baker, 2016), 70% of researchers reported failing to replicate another's work and more than half reported failing to reproduce their own studies. This extraordinary lack of replication is alarming in its own right and points to questionable research practices. Many respondents from the *Nature* survey stated that the crisis could be solved by 'More robust experimental design', 'better statistics' and 'better mentorship' (Baker, 2016). Journals can assist with all of these challenges.

Journals are part of the life-cycle of research and as such we have a responsibility to sustain and improve scientific quality in the papers we publish. In the field of psychology, most attention has been directed to social and cognitive psychology, but the issues have relevance to clinical psychology as well (Tackett et al., 2017). The goal of this special issue is to discuss issues surrounding research practices and reproducibility within behavioral research and therapy, the challenges we face in addressing these issues, and the steps *Behaviour Research and Therapy* is taking to improve the rigor and integrity of our reporting and protect our science from biases.

Our introductory piece overviews the issues and steps that *Behaviour Research and Therapy* is taking to reduce biases, increase the standards of methods and reporting and increase reproducibility. Waters, LeBeau, Young, Dowell and Ryan present a set of recommendations for improving the quality of research practices, specifically with a focus upon the quality of published manscripts. They present a framework for maintaining quality in the publication lifecyle. Furthermore, they provide suggestions for ways in which funding bodies, research institutions, journals and peer reviewers can encourge and support quality publication processes (QPP). Hildebrandt and Prenoveau outline the controversies within statistical sciences that threaten rigor and reproducibility of research published in our field and discuss ongoing approaches to generate reliable and valid inferences from data. Among the major issues are methods for establishing hypotheses before data analyses are conducted, misuse of statistical inference, and the prevalence and drawbacks of p-hacking approaches. They also discuss emerging approaches for peer review to achieve these goals and make recommendations for *Behaviour Research and Therapy* reviewers along these lines. Cristea and Naudet present issues around the topic of research waste within the context of psychological treatment research. They examine waste in terms of research priorities, research design and methods, accessibility of information and accuracy of publications. They also suggest possible solutions. Dunn and colleagues address the tensions between approaches designed to reduce research waste by improving research rigour, on the one hand, and scientific innovation on the other hand. They make their case with the example of depression treatment research, but also conclude that lack of rigor is only one cause of research waste and that attention to research priorities would more effectively reduce waste.

1. State of the field

Biomedical research in general has been criticized for poor quality, with estimates as high as 50% of published research reports being sufficiently poor as to make them unusable and a waste (Chalmers & Glasziou, 2009). The poor quality derives from both reporting deficiencies as well as deficiencies in original design or methodology. With respect to reporting deficiencies, adequate information about biomedical interventions is estimated to be available in about only 65% of cases, with an even lower rate in systematic reviews (Glasziou, Meats, Heneghan & Shepperd, 2008). The use of guidelines such as CONSORT (Consolidated Standards of Reporting Trials) have improved reporting but adherence remains problematic, such as inadequate descriptions of sample characteristics and the flow of participants from recruitment through allocation, differences between analyses and outcomes in the published report versus the original study protocol, and limited to no information about adverse effects (Glasziou et al., 2014). For example, one review indicated that primary outcomes were changed, introduced or omitted in 62% of reports (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004).

Reporting violations are closely related to a questionable research practice referred to as p-hacking, or an attempt to observe a p value below the threshold for significance, which obviously leads to biases and increases in Type I error (Tackett et al., 2017). In their influential article, Simmons, Nelson, and Simonsohn (2011) presented six ways to p-hack: stop collecting data once $p < .05$; analyze many measures but report only those with $p < .05$; collect and analyze many conditions but only report those with $p < .05$; use covariates to get $p < .05$; exclude participants to get $p < .05$; and transform the data to get $p < .05$. Not all p-hacking is bad: p-hacking practices may be suitable in certain exploratory contexts, when recognized as such, but they are problematic when reported as if they were predicted when in fact they were exploratory. A related reporting violation is coined HARKing (Kerr, 1998) or presenting hypotheses as a priori when in fact they

https://doi.org/10.1016/j.brat.2019.03.006

Available online 06 March 2019
0005-7967/

developed once the results were known.

Limitations in original design or methodology include factors such as limited sample sizes and related power, poor measurement and variable quantification, and inadequate comparison or control groups (e.g., Tackett et al., 2017). Our field of clinical science has been particularly hampered by underpowered studies, if only due to limited funding and the intensive effort required to recruit and study samples of individuals at risk for or experiencing mental health problems. Coyne and colleagues have commented very critically upon these issues, particularly as they apply to health psychology (Coyne, 2016). Given this state of affairs, Coyne argues that we "stop allowing small studies from entering effect sizes into the literature and withhold judgments about efficacy until the availability of larger-scale, more methodologically sophisticated studies" (2016, p. 4–5).

The combination of poor reporting and limited design or methodology most likely contributes to the high rates of irreproducibility. Although clinical science has not been the focus of the replicability crisis, we should be concerned, especially given the potentially high impact of our findings upon the delivery of mental health services (Tackett et al., 2017). As Tackett et al. (2017) point out, our current standard for recognizing an intervention as empirically supported relies upon only two rigorous and independently conducted randomized controlled trials, and does not take into account the number of negative results. They further note that we need to think carefully about more rigorous standards for judging an intervention to be empirically supported since such a 'standing' essentially gives the go-ahead for the intervention be implemented widely and affect many lives.

The responsibility for scientific rigor is a community issue. This community includes not just the researcher but institutes, funders, journals, industry and scientific societies, and that is just naming the main players. There is a larger system of incentives at play that influence practices both explicitly and implicitly. Most pervasive among these are pressures to publish in order to advance, receive funding or praise. "A focus on publication of reports in journals with high impact factors and success in securing funds leads scientists to seek short-term success instead of cautious, deliberative, robust research that will take substantially longer to produce less exciting (but more valid) findings" (Macleod, Michie, Roberts, Dirnagl, Chalmers, Ioannidis, Salman, Chan & Glasziou, 2014, p. 103). Institutes reward high-profile publications, both in terms of money and career progression, and funders focus on novelty. Researchers are not incentivized to do the 'groundwork' studies, replicate their own research or others. Journals have also been accused of only wanting to publish 'sexy' new research rather than replications, negative findings, and so on. The concept of 'normal science' put forward by Thomas Kuhn in his book 'The Structure of Scientific Revolutions' (Kuhn, 1962, p. 1962) has been passed over for the 'ground breaking', undermining the role 'normal science' plays in the accumulation of knowledge which eventually leads to paradigm shifts. We are in need of a shift in the incentive structure, such that decisions regarding promotion, funding and publication are based on qualities of methodological rigour, open data sources, quality of reporting and reproducibility (Macleod et al., 2014). For our part, *Behaviour Research and Therapy* is considering reviewing and publishing protocols with guarantee of publishing results regardless of the outcomes, consistent with the article-type Registered Reports method described below, and requesting a series of replications on selected innovative findings.

2. Potential solutions and challenges for Behaviour Research and therapy

In addition to shifts in system-wide incentive structures, we need to address errors or biases that occur in the reporting of research and the limitations in original methods and design. In terms of reporting, solutions that have been offered include (1) standards of reporting such as Consolidated Standards of Reporting of Trials (CONSORT) (2) pre-registration of rationale, design, analytic plans, and primary outcomes and (3) sharing of data. Each of these solutions poses challenges, such as the challenge of implementation for reporting standards. To address this challenge, some journals have created their own checklists to facilitate standards of reporting; *Cell* has created the STAR (Structured, Transparent, Accessible Reporting) methods (https://www.cell.com/star-authors-guide) and *Nature* has a reporting summary checklist (https://www.nature.com/authors/policies/ReportingSummary.pdf), both of which aim to improve the quality of reporting. The most important aspect of any standard and the key to their success is their adoption by the community. Authors can sometimes perceive these standards as bureaucracy, 'another hoop to jump through' to get papers published. The community has to sign-up to the standards, agreeing that they will improve the quality of science and that they must be considered at study initiation and not after the data have been collected and analyzed. Adoption of these standards is not restricted to the authors but also the reviewers and editors who will have to spend additional time checking adherence. Ways of encouraging *Behaviour Research and Therapy* reviewers to check adherence to reporting standards is an area we are investigating. One example is the Peer Reviewers' Openness Initiative, in which peer reviewers refuse to conduct a review of a research paper until and unless related materials are either made available or a good reason for why that is not possible is provided. However, there are risks associated with this particular initiative, at least as recognized by Bishop (2016).

Pre-registration is an important deterrent to p-hacking and HARKing, and is a condition that *Behaviour Research and Therapy* has required since October 2016. Pre-registration can be completed on the Open Science Platform (osf.io) or the National Institutes of Health site (clinicaltrials.gov). Yet, challenges exist, and, as reviewed by Tackett et al. (2017), clinical science faces very specific challenges in this regard. For example, clinical realities (such as closure of a clinical site for recruitment) sometimes require changes to design or to methods midstream (Tackett et al., 2017). Flexible pre-registration schemes provide one solution to necessary modifications while simultaneously addressing another issue, which is the impediment to scientific discovery that comes with strict adherence to pre-registered hypotheses. Rigidity in adherence prohibits an "Iterative process between theory and data [as] a fundamental component of a growing science" (Tackett et al., 2017, p. 749). Others similarly note that "scientific progress also depends on serendipitous discoveries from open-ended exploration. Exploratory analyses are to be encouraged, provided it is recognized that their use is for generating hypotheses that should then be tested in a new data set" (Bishop, 2016, p. 5). In this special issue, Dunn and colleagues discuss the tension between standards for reporting and pre-registration on the one hand, and innovation in science on the other hand. Flexibility in pre-registration that allows hypotheses and plans to be modified seems advisable, as long as each change is accompanied by a time stamp and explicit statement to indicate how and why the change is made.

Aside from pre-registration, journals can permit article-type Registered Reports, as was launched by the journal *Cortex*. In this article-type, the methods and proposed analyses are pre-registered and peer-reviewed prior to research being conducted. The protocols are provisionally accepted by the journal and once the study has been completed, reviewers again appraise the paper to ensure the required conditions have been met (Chambers, 2013). Authors are free to submit their full study to a journal other than the one in which they registered their study, but under those conditions, the original journal will publicly record the registration in a section called Withdrawn Registrations. So far, uptake of this article type has been low (although still in its infancy) and there are issues with ensuring fair peer review as the time between registration and completion of the study can be several years, such that the same reviewers may not be available to assess the completed study.

Data availability and open science is another recommendation for reducing questionable practices and improving reproducibility. The Open Science Framework provides a platform on which to not only pre-

register but also share data and data analysis. Elsevier has a journal data-sharing policy that is aligned with the data guidelines that are part of the Transparency and Openness Promotion (TOP) guidelines as defined by the Center for Open Science (https://www.elsevier.com/journals/behaviour-research-and-therapy/0005-7967/guide-for-authors), and *Behaviour Research and Therapy* has endorsed these guidelines. Specifically, Elsevier 'encourages and supports researchers to share research data where appropriate and at the earliest opportunity' (see https://www.elsevier.com/about/policies/research-data for the full policy) and offers authors a variety of ways to share their data in their article. These include data linking in their papers, Mendeley Data to deposit data and 'data statements' to state the availability of the data. Several challenges to data sharing exist within the field of clinical science. First, the extensive efforts required, where "recruiting clinical participants can be considerably more difficult, time consuming, and resource-intensive than using healthy or convenience samples" (Tackett et al., 2017, p. 751), can drive investigators away from sharing of data. On the other hand, we fully endorse the counter-argument that open science increases the value of research data and that as scientists were are ethically responsible to maximize use of data that is publicly funded (Bishop, 2016). A more concerning barrier is the sensitivity of the data collected (e.g., traumatic life events, diagnoses) which may be difficult to fully de-identify, especially in small or specialized samples (Tackett et al., 2017). Moreover, secondary analysts may utilize the data to assess certain sensitive issues for which participants had not originally consented (Bishop, 2016). Paradoxically, access to large data sets that are already collected may also increase the risk of p-hacking, although pre-registration for secondary data analysis can address this concern (Bishop, 2016).

Another potential solution is "an independent statistics and methods unit that could take a fresh look at existing data on any topic where there is substantial public concern that the findings may be flawed" (Bishop, 2016, p. 6). An independent statistical review offers many advantages, although faces many challenges in implementation, such as cost. In response to this concern, we instituted a Statistical Editorial Board for *Behaviour Research and Therapy*, so that editors could consult or request reviews from statistical experts whenever they perceived the need. Nonetheless, challenges remain, particularly when reviewing meta-analyses (for which methodological details are even less well reported than in empirical studies (Glasziou, Meats, Heneghan & Shepperd, 2008) and which are rapidly growing in number), as the effort required to conduct a full independent statistical review of meta-analyses is beyond the scope of either reviewers or editorial board members. One potential option is to require all submissions of meta-analyses to be accompanied by author-assurance of an independent statistical review, although of course this imposes costs upon the researcher. Conceivably, funding agencies could support costs of independent statistical review if designed in advance. Another solution is for authors of meta-analysis to post their data in an open repository, something that is possible within our data sharing platform for *Behaviour Research and Therapy*. Very importantly, institutes need to assess how they continually educate their researchers on these issues.

Independent replication is important for scientific standing and verification. However, replication is a particular challenge in clinical science, due to factors such as recruitment constraints (e.g., low base rate conditions, sampling variations across clinical populations, and even changes in diagnostic criteria) and again the effort involved in intensive data collection may deter investigators from establishing replication, especially when systems continue to value novel findings (Tackett et al., 2017). Tackett et al. (2017) recommend that more energy be devoted to independent replications via access to original data sets (as part of the open science agenda). Another suggestion they make is to build replication into study design. Unfortunately, there is frequently neither the time nor the resources to replicate studies and following in another researcher's footsteps is not seen as an optimal way for career advancement. These issues disincentivize replications, and as described earlier, the system-wide incentive structure must be revised to give more value to 'normal science'.

Questionable research practices and ultimately replicability can be targeted through improved study method and design. As Tackett et al. (2017) recommend in the context of clinical science, this includes assurance of adequate sample size (with power analyses in advance of data collection), greater measurement precision, steps to reduce error in variable quantification (such as by use of multiple measures of key constructs and aggregation across measures to form latent variables) and increased use of within subject research. Complete step-by-step protocols can be registered in Protocols.io (https://www.protocols.io/) and then linked in the Methods sections of published papers to increase transparency in methods. Furthermore, increased harmonization of measures across laboratories would allow pooling of data to increase sample sizes. Such efforts are already underway, such as the NIH Toolbox (www.healthmeasures.net). The tensions between commonality in measurements on the one hand and unique or innovative measures on the other is discussed by Dunn and colleagues in this special issue.

3. Conclusion

Science, at its core, is about knowledge accumulation but, as our understanding increases, so do our methods and standards and we have to adapt to increase our standards as necessary. There is no end point in knowledge, but we have to ensure that the rigour 'road' is as robust as we can make it. Science is always a 'work in progress' and consequently so will be our protocols and standards.

Acknowledgements

I wish to thank Kate Wilson for her thorough reading and extensive comments upon this manuscript.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452–454. https://doi.org/10.1038/533452a.

Bishop, D. V. M. (2016). Open research practices: Unintended consequences and suggestions for averting them. (Commentary on the peer reviewers' openness initiative). *Royal Society Open Science, 3*(4), 160109. https://doi.org/10.1098/rsos.160109.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet, 374*(9683), 86–89. https://doi.org/10.1016/s0140-6736(09)60329-9.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex [editorial]. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016.

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials. *Journal of the American Medical Association, 291*(20), 2457. https://doi.org/10.1001/jama.291.20.2457.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4*(1)https://doi.org/10.1186/s40359-016-0134-3.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet, 383*(9913), 267–276. https://doi.org/10.1016/s0140-6736(13)62228-x.

Glasziou, P., Meats, E., Heneghan, C., & Shepperd, S. (2008). What is missing from descriptions of treatment in trials and reviews? *BMJ, 336*(7659), 1472–1474. https://doi.org/10.1136/bmj.39590.732037.47.

Ioannidis, J. P. A. (2016). Why most clinical research is not useful. *PLoS Medicine, 13*(6), e1002049. https://doi.org/10.1371/journal.pmed.1002049.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Kuhn, T. S. (1962). *The structure of scientific Revolutions.* Chicago: University of Chicago Press0-226-45808-3.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., ...

Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet, 383*(9912), 101–104. https://doi.org/10.1016/s0140-6736(13)62329-6.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., ... Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12*(5), 742–756. https://doi.org/10.1177/1745691617690042.

Michelle G. Craske
Department of Psychology, University of California, Los Angeles, United States
E-mail address: mcraske@mednet.ucla.edu.