Chapter 10
Open Science, Replicability, and Transparency in Modelling

Toby Prike

Recent years have seen large changes to research practices within psychology and a variety of other empirical fields in response to the discovery (or rediscovery) of the pervasiveness and potential impact of questionable research practices, coupled with well-publicised failures to replicate published findings. In response to this, and as part of a broader open science movement, a variety of changes to research practice have started to be implemented, such as publicly sharing data, analysis code, and study materials, as well as the preregistration of research questions, study designs, and analysis plans. This chapter outlines the relevance and applicability of these issues to computational modelling, highlighting the importance of good research practices for modelling endeavours, as well as the potential of provenance modelling standards, such as PROV, to help discover and minimise the extent to which modelling is impacted by unreliable research findings from other disciplines.

10.1 The Replication Crisis and Questionable Research Practices

Over the past decade many scientific fields, perhaps most notably psychology, have undergone considerable reflection and change to address serious concerns and shortcomings in their research practices. This chapter focuses on psychology because it is the field most closely associated with the replication crisis and therefore also the field in which the most research and examination has been conducted (Nelson et al., 2018; Schimmack, 2020; Shrout & Rodgers, 2018). However, the issues discussed are not restricted entirely to psychology, with clear evidence that similar issues can be found in many scientific fields. These include closely related fields such as experimental economics (Camerer et al., 2016) and the social sciences more broadly (Camerer et al., 2018), as well as more distant fields such as biomedical research (Begley & Ioannidis, 2015), computational modelling (Miłkowski et al., 2018), cancer biology (Nosek & Errington, 2017), microbiome research

J. Bijak, *Towards Bayesian Model-Based Demography*, Methodos Series 17,
https://doi.org/10.1007/978-3-030-83039-7_10

(Schloss, 2018), ecology and evolution (Fraser et al., 2018), and even within methodological research (Boulesteix et al., 2020). Indeed, many of the lessons learned from the crisis within psychology and the subsequent periods of reflection and reform of methodological and statistical practices apply to a broad range of scientific fields. Therefore, while examining the issues with methodological and statistical practices in psychology, it may also be useful to consider the extent to which these practices are prevalent within other research fields with which the modeller is familiar, as well as the research fields that the findings of the modelling exercise either relies on, or is applied to.

Although there was already a long history of concerns being raised about the statistical and methodological practices within psychology (Cohen, 1962; Sterling, 1959), a succession of papers in the early 2010s brought these issues to the fore and raised awareness and concern to a point where the situation could no longer be ignored. For many within psychology, the impetus that kicked off the replication crisis was the publication of an article by Bem (2011) entitled "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect." Within this paper, Bem reported nine experiments, with a cumulative sample size of more than 1000 participants and statistically significant results in eight of the nine studies, supporting the existence of paranormal phenomena. This placed researchers in the position of having to believe either that Bem had provided considerable evidence in favour of anomalous phenomena that were inconsistent with the rest of the prevailing scientific understanding of the universe, or that there were serious issues and flaws in the psychological research practices used to produce the findings.

Further issues were highlighted through the publication of two studies on questionable research practices in psychology, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant" by Simmons et al. (2011), and "Measuring the prevalence of questionable research practices with incentives for truth telling", by John et al. (2012). Using two example experiments and a series of simulations, Simmons et al. (2011) demonstrated how a combination of questionable research practices could lead to false-positive rates of 60% or higher, far higher than the 5% maximum false-positive rate implied by the endorsement of $p < 0.05$ as the standard threshold for statistical significance. Specifically, the authors showed that collecting multiple dependent variables, not specifying the number of participants in advance, controlling for gender or the interaction of gender with treatment, or having three conditions but preferentially choosing to report either all three or only two of the conditions, can lead to large increases in the false-positive rates that become even more extreme when several of these research practices are combined. To drive home the point further, Simmons et al. (2011) conducted a real study with 20 undergraduate students and then used the analytical flexibility available to them and the lax reporting standards for statistical analyses to report an impossible finding: that they had 'found' that listening to the song "When I'm Sixty-Four" rather than "Kalimba" led to participants being younger, with the test statistic $F(1, 17) = 4.92$ implying a 'significant' p-value, $p = 0.040$.

Closely following the Simmons et al. (2011) paper, John et al. (2012) published a survey on the research practices of psychologists, finding that the type of practices Simmons et al. (2011) had shown to be highly problematic were commonplace. Responses to the full list of questionable research practices included in the survey varied considerably (see John et al., 2012 for full results for all ten questionable research practices). Some research practices were considered much less defensible, such as outright falsification of data (admitted to by 0.6–1.7% of the sample of researchers, depending on the condition) or making misleading or untrue statements within the paper such as, "In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)", (admitted to by 3.0–4.5% of the sample, depending on condition). Even more commonplace was the benefit of hindsight: the statement, "In a paper, reporting an unexpected finding as having been predicted from the start", was admitted to by 27.0–35.0% of the sample, again depending on condition (John et al., 2012, *passim*).

Other research practices examined in the survey were considered more defensible and were admitted to by a majority of the psychologists surveyed, but can still contribute to massively increased false positive rates prevalent in the literature. For example, 55.9–58.0% of the sample admitted to, "Deciding whether to collect more data after looking to see whether the results were significant", and 63.4–66.5% of the sample admitted to, "In a paper, failing to report all of a study's dependent measures" (*idem*). It is also important to note that these are conservative estimates based on the willingness of individual psychologists to admit that they personally had engaged in questionable research practices, and therefore the actual prevalence of questionable research practices is likely far higher. John et al. (2012) also calculated prevalence estimates based on respondents' answers to questions about the percentage of *other psychologists who have engaged in a questionable research practice* as well as the percentage of those *other psychologists who have engaged in a questionable research practice and would admit to having done so,* and for nearly all of the questionable research practices these estimates were considerably higher than the number who actually made self-admissions within the survey (*idem*).

The publication of a large-scale replication attempt of 100 psychological findings by the Open Science Collaboration (2015) showed the practical extent of the problems highlighted by Simmons et al. (2011) and John et al. (2012). Although 97 of the 100 original studies included for replication reported statistically significant results, only 36 of the replication attempts ended up statistically significant, despite having statistically well-powered designs (with an average power – probability of correctly rejecting a false hypothesis – equal to 0.92), and despite matching the original studies closely, including using original materials wherever possible. Other large-scale replication efforts, including the Many Labs projects within psychology (Ebersole et al., 2016; Klein et al., 2014, 2018), projects in fields such as experimental economics (Camerer et al., 2016), and the social sciences more broadly (Camerer et al., 2018), as well as more distant fields, such as cancer biology (Nosek & Errington, 2017), have highlighted that, to varying extents, there are serious issues with the reliability and replicability of findings published within many scientific areas.

10.2 Open Science and Improving Research Practices

Once the issues outlined above were clearly highlighted, many scholars within psychology decided that reform was necessary, and serious changes within the field needed to be made.[1] Changes to current practices were recommended at several levels of the scientific process, including at the level of individual authors, reviewers, publishers, and funders (Munafò et al., 2017; Nosek et al., 2015; Simmons et al., 2011). Some of the changes to research practice that have been most commonly recommended and widely engaged with by researchers include openly publishing the data and analysis code online, openly publishing study materials online, and the preregistration of study methodology and analysis plans (Christensen et al., 2019).

The change in research practice that has seen the earliest and greatest uptake by researchers is the public sharing of data and/or analysis code (Christensen et al., 2019). Making the data and analysis code underlying research claims openly available has many potential benefits for both science as a whole and for individual researchers who engage in the practice. Benefits to the scientific process from the open sharing of data include: allowing other scientists to re-analyse data to help verify the results and check for errors, providing safeguards against misconduct such as data fabrication, or taking advantage of analytical flexibility, for example, because other scientists can discover that a result is entirely reliant on a specific covariate. It also allows other researchers to reuse the data for a variety of purposes (Tenopir et al., 2011). If data are publicly available, then they may be reanalysed to answer new questions that were not initially examined by the researchers. Without open data, these reanalyses would not be possible and therefore the scientific knowledge would either not be generated at all, or would require the recollection of the same, or highly similar data, leading to waste and inefficiency in the use of resources (usually public funding; Tenopir et al., 2011).

There are also good reasons for individual researchers to publicly post their data even if they are motivated by their own self-interest. Articles with publicly available data have an advantage in the number of citations received (Christensen et al., 2019; Piwowar & Vision, 2013), and willingness to share data are associated with the strength of evidence and quality of the reporting of statistical results (Wicherts et al., 2011). However, even though the uptake of the public posting of data and software code is growing quickly and should be lauded, there are still many problematic areas, such as incomplete data, missing instructions, and insufficient information provided. These issues mean that even when data are publicly shared, independent researchers may still regularly face considerable hurdles and/or not actually be able to analytically reproduce the results reported in the paper (Hardwicke et al., 2018; Obels et al., 2020; Stagge et al., 2019; Wang et al., 2016).

[1] Although it has to be noted that there was also pushback from some scholars – see Schimmack (2020) for further discussion of the responses to the replication crisis.

Another common and rapidly growing area of open science is the public posting of study materials or instruments and experimental procedures (Christensen et al., 2019). Like open data and analysis code, this practice has the benefit of increasing transparency and making it clear to editors, reviewers, and readers of articles, what exactly was done within the study. This increased transparency allows for easier assessment of whether there are potential confounds or other flaws in the study methodology that may have impacted on the conclusions. It also allows for easier assessment of the appropriateness and validity of the stimuli and materials used. Openly sharing materials and procedures also has the additional benefits of making it far easier for other researchers to conduct direct *replications* of the research (i.e., taking the same materials and procedures and collecting new data to independently verify the results), as well as to conduct follow up studies that attempt to conceptually replicate, adapt, or expand on some or all of the aspects of the study without the need to contact the original authors and/or to expend time and resources reproducing or creating new study materials and procedures. These practices are in addition to ensuring the *reproducibility* of the results, which is here understood as ensuring that the software or computer code applied to a given dataset produces the same set of results as reported in the study.[2]

One major change in research practice that has the potential to greatly reduce questionable research practices and improve the quality of science is preregistration: registering the aims, methods and hypotheses of a study with an independent information custodian *before* data collection takes place (Nosek et al., 2018; Wagenmakers et al., 2012). Although preregistration is still currently less common than openly sharing data, code, and materials, the uptake of the practice is increasing rapidly (Christensen et al., 2019). Preregistration has been referred to as 'the cure' for analytical flexibility or 'p-hacking', the practice of fine-tuning analyses until the desired or a publishable result, as measured by the magnitude of p-values, can be obtained (Nelson et al., 2018, p. 519).

When researchers preregister their studies, they need to outline in advance what their research questions and hypotheses are, as well as their plans for analysing the data to answer these questions and verify the hypotheses (Nosek et al., 2018; Wagenmakers et al., 2012). Therefore, if done correctly, preregistration ensures that the analyses conducted are confirmatory, which is a required assumption for null hypothesis significance testing. It also allows both the researchers themselves and other consumers of research products to have much greater confidence that the results can be relied upon, and the false-positive rate has not been greatly inflated through questionable research practices (Simmons et al., 2011). In this way, preregistration is also useful for the researchers conducting the research, as it helps them to avoid biases and misleading themselves (Nosek et al., 2018). Once discovering an unexpected but impactful result in the data, or that controlling for a variable or excluding participants based on a specific criterion leads to a statistically significant

[2] For a broad terminological discussion of replicability and reproducibility, which are terms that still remain far from being unambiguously defined and used, see e.g. National Academies of Sciences, Engineering, and Medicine (2019).

finding that can be published, it can be easy for hindsight bias and wishful thinking to lead researchers to justify these analytical decisions to both themselves and others, and to believe that they predicted or planned them all along (also known as 'hark-ing' – "hypothesising after results are known"; Kerr, 1998).

However, preregistration alone is not likely to solve the problems with research malpractice unless reviewers, editors, publishers, and readers ensure that researchers actually follow their preregistered hypotheses and analysis plans. Registration of clinical trials has been commonplace for some time now, yet published trials still regularly diverge from the prespecified registrations, with publications switching and/or not reporting the primary outcomes listed in trial registries (Goldacre et al., 2019; Jones et al., 2015), and journals showing resistance to attempts to highlight or correct issues when informed of discrepancies between the trial registries and the articles they had published (Goldacre et al., 2019). Going even further than preregistration, a growing number of journals now offer a registered report format in which studies are reviewed based on the underlying research question(s), study design, and analysis plan and can then be given in principle acceptance, meaning that the study will be published regardless of the results provided the authors adhere to the pre-agreed protocols (Chambers 2013, 2019; Nosek & Lakens, 2014; Simons et al., 2014).

In addition to the changes in research practice outlined above, there has also been considerable discussion about the use of statistics within psychology and other scientific fields, including a special issue of *The American Statistician* entitled "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$". Within the special issue, and in various other articles, books, and publications, the contributors have criticised the use of p-values, and particularly the $p < 0.05$ cut-off conventionally used to determine 'statistical significance', as well as the phrase 'statistically significant' itself. Indeed, the editors of *The American Statistician* recommended that the phrase 'statistically significant' no longer be used (Wasserstein et al., 2019).

There is still much disagreement about what new statistical practices should be adopted or how researchers should move forward, with a variety of potential solutions proposed. For example, some have recommended that the $p < 0.05$ threshold be redefined to $p < 0.005$ instead (Benjamin et al., 2018), whereas others have advocated for a shift away from null hypothesis significance testing towards Bayesian analyses and inference (Wagenmakers et al., 2018). At the same time, some other authors, notably Gigerenzer and Marewski (2015), have warned about the idolisation of simple Bayesian measures, such as Bayes Factors. In the same way as had happened with p-values, indolent statistical reporting can occur under the Bayesian paradigm as much as in the frequentist one. Although there is still some disagreement about the possible future directions for statistical analysis and inference, the general guidance provided by the editors of *The American Statistician* – "Accept uncertainty. Be thoughtful, open, and modest." (Wasserstein et al., 2019, p. 2) – provides a direction for future empirical enquiries.

10.3 Implications for Modellers

The above discussion has outlined a series of issues that have occurred within psychology and a variety of other experimental and empirical domains of science, as well as some of the solutions that are already being implemented and potential future directions for further improvements in methodology and statistics. The following section relates these considerations back to the specific domains of computational modelling and simulation, highlighting the relevance of the lessons learned for researchers and practitioners within these domains. There is documented evidence of similar issues occurring within computational modelling, and issues within empirical fields can also impact computation modelling because of the interconnectedness of scientific disciplines.

Many of the issues highlighted above are also relevant for computational modelling, and even in circumstances where a concern is not directly applicable to modelling challenges, there are some analogous concerns (Miłkowski et al., 2018; Stodden et al., 2013). As with the practice of sharing data, analysis code, study materials, and study procedures for empirical studies, clearly and transparently documenting models is vital for other researchers to be able to verify and expand upon existing work. Chapter 7 of this book highlights several existing methods that modellers can use to document or describe simulation models, such as the ODD protocol (Overview, Design concepts, Details; Grimm et al., 2006), or provenance standards, such as PROV (Groth & Moreau, 2013).

Similar to the sharing of data and analysis code, there are often serious issues with attempting to computationally reproduce existing models and simulations even if code is provided. This can happen because of a range of factors, such as the exclusion of important information within publications and failing to properly document model and/or simulation code (Miłkowski et al., 2018). As with sharing data and analysis code for empirical work, transparently sharing documentation and descriptions of computational models has the advantage of allowing other researchers to test and verify the extent to which outputs are dependent on specific modelling choices made in the modelling process, how sensitive the model is to changes in various inputs (see Chap. 5 for more details on sensitivity analysis), and/or the extent to which the results change (or remain consistent) when the model uses different data or is applied in a different context (e.g., if a model of asylum migration from Syria is applied to asylum migration from Afghanistan).

Computational modelling often requires far more decisions regarding design, formalisation, and implementation than standard experimental or empirical work, and in some cases is more exploratory in nature. Therefore, preregistration does not seem like a readily applicable or appropriate format to be transferred to all aspects of computational modelling, although it is certainly still applicable to at least some aspects (e.g., if models are to be compared, it is useful to preregister the models that will be compared as well as how the comparison will be conducted; see Lee et al., 2019 for more information). Nonetheless, there are several strategies that can be used to try and reduce the extent to which modellers have the flexibility to tinker with their models to find the specific settings that produce the desired (publishable) results.

One option here is for modellers to develop and rely on prespecified architectures within their models, such as the BEN (Behavior with Emotions and Norms) architecture, which provides modules that can add aspects such as emotions, personality, and social relationships to agent-based models (Bourgais et al., 2020). Alternatively, independent researchers can recreate a model without referring to or relying on the original model code, which can help to test the extent to which outputs are dependent on modelling choices for which there are a variety of plausible and defensible alternative options (see Silberzahn et al., 2018 for an analogous example with statistical analyses). Reinhardt et al. (2019) have provided a detailed discussion of the processes and lessons learned from implementing the same model in two different modelling languages, one a general-purpose language using discrete-time and the other a domain-specific modelling language using continuous time.

In addition to the open science and methodological concerns within computational modelling, related research practices within psychology and other empirical fields can also have considerable impact on modelling practice because of the interplay between scientific disciplines and how computational models may rely on or be informed by findings from empirical work. Therefore, the tendency for many empirical fields to simply rely on finding ‘statistically significant’ effects rather than attempt to accurately estimate effect sizes or test them for robustness limits the extent to which these findings can be usefully and easily applied to computational models. Additionally, if a computational model is informed by, or relies on, empirical findings to justify mechanisms and processes within the model (e.g., the decision making of agents within an agent-based model), then if those findings are unreliable and/or based on questionable research practices, this may effectively undermine the whole model.

These limitations once again highlight the advantage of provenance modelling standards, such as PROV (Groth & Moreau, 2013; Ruscheinski & Uhrmacher, 2017), as a format for documenting and describing models. PROV allows information to be stored in a structured format that can be queried, thereby allowing it to be easily seen which entities a model relies on (see Chap. 7). Therefore, if new research highlights issues within the existing literature (e.g., a failed replication within psychology), or new discoveries are made, it is a relatively simple and straightforward task to search PROV information, and discover which models have incorporated this information as an entity, and therefore may have at least some aspects of the model that need to be reconsidered or updated.

This strategy could also be combined with sensitivity analysis (see Chap. 5) to establish the extent to which the model outputs are sensitive to aspects that rely on the entity now called into question, and therefore whether it is necessary to update the model in light of the new information. Additionally, PROV has the potential to contribute to the empirical literature by highlighting specific entities (e.g., research studies) that are commonly featured within models. Such studies may therefore become a high priority for large-scale replication efforts, not only to ensure the reliability and robustness of the findings, but also to identify potential moderators (mediating and confounding variables) and boundary conditions.

The choice of specific tools and solutions notwithstanding, one lesson for modellers that can be learned from the replicability crisis is clear: transparency and proper documentation of the different stages of the modelling process are vital for generating trust in the modelling endeavours and in the results that the models generate. For the results to be scientifically valid, they need to be reproducible and replicable in the broadest possible sense – and documenting the provenance of models is a necessary step in the right direction.