Witold. M. Hensel

Double Trouble? The Communication Dimension of the Reproducibility Crisis in Experimental Psychology and Neuroscience[*†]

Abstract

Most discussions of the reproducibility crisis focus on its epistemic aspect: the fact that the scientific community fails to follow some norms of scientific investigation, which leads to high rates of irreproducibility via a high rate of false positive findings. The purpose of this paper is to argue that there is a heretofore underappreciated and understudied dimension to the reproducibility crisis in experimental psychology and neuroscience that may prove to be at least as important as the epistemic dimension. This is the communication dimension. The link between communication and reproducibility is immediate: independent investigators would not be able to recreate an experiment whose design or implementation were inadequately described. I exploit evidence of a replicability and reproducibility crisis in computational science, as well as research into quality of reporting to support the claim that a widespread failure to adhere to reporting standards, especially the norm of descriptive completeness, is an important contributing factor in the current reproducibility crisis in experimental psychology and neuroscience.

Introduction

The reproducibility crisis is usually depicted as resulting from a failure to follow some norms of scientific investigation (but see, e.g., Bird 2018). If sufficiently widespread, this failure can cause a proliferation of false positive findings in the literature, which can lead to low research reproducibility because false findings are mostly irreproducible. For example, many authors have argued that researchers in psychology, including journal editors and referees, tend to confuse nominal statistical significance of an experimental finding with its epistemic import (e.g., Rosenthal & Gaito 1963; Neuliep & Crandall 1990, 1993; Sterling 1959; Sterling, Rosenbaum, & Weinkam 1995; Gigerenzer 2004, 2018). In effect, the psychological literature has been filling up over the years with statistically significant results, most of which are probably false. The problem is serious as most psychology papers use null hypothesis significance testing (NHST) and over 90% of them report positive findings (Fanelli 2010).

However, besides the norms of scientific investigation, there are norms of scientific reporting, whose violation can also affect reproducibility even if it does not necessarily lead to a proliferation of false positives in the literature. Chief among these norms is the norm of descriptive completeness, which states that a research report should provide all the information necessary to enable a knowledgeable independent researcher to recreate the study. Widespread failure to follow the norm of descriptive completeness would have disastrous consequences, [illegible]redibility.

Yet scientists and philosophers of science seldom pay attention to the norms of scientific reporting. We take it for granted that researchers have no difficulty producing adequate descriptions of their experiments. After all, designing and conducting a scientific study is much more intellectually demanding than producing a research

---

[*] Forthcoming in the *European Journal for Philosophy of Science*.

[†] I would like to express my gratitude to Maja Białek, Mateusz Hohol and Marcin Miłkowski for many valuable discussions on the reproducibility crisis. I would also like to thank Bartosz Maćkiewicz and the participants of a philosophy of cognitive science seminar conducted by Marcin Miłkowski at the Polish Academy of Sciences for helpful comments on a first draft of this paper. Last but not least, I thank the anonymous reviewers of this journal for constructive criticism.

report. Any high school graduate can write a report. And, let's face it, the norms of scientific reporting are not as interesting as the norms of scientific investigation. The most general ones, which apply to any report regardless of study design or field, are pretty much self-evident. We all agree that a research report should be clear, concise, true and complete. What is more problematic is how to apply these norms in specific cases. To take the norm of descriptive completeness again, it is not obvious which details of, say, randomized controlled trials are essential to reproducibility. Which sampling decisions need to be reported and where – in the main text, in supplementary materials, in a repository? And what about information such as room temperature or the weather outside? Although nontrivial, such questions are frequently dismissed as pedantic and irrelevant. We often feel that only a pedant would seek precise answers to them. A reasonable person would simply err on the side of caution, wouldn't they?

The aim of this paper is to show that the pedants may actually be on to something. I contend that the norms of scientific reporting are violated much more frequently than we like to admit. More specifically, I argue that violations of the norm of descriptive completeness are probably so widespread as to affect research reproducibility in experimental psychology and neuroscience. Indeed, I claim that, without recognizing this, we will not be able to properly assess either the nature or the true scale of the reproducibility crisis. Although the open science movement advocates transparency and clear communication, empirical assessments of quality of reporting in experimental psychology and neuroscience are few and far in between. This makes it impossible to estimate the actual impact of reporting practices on reproducibility with any confidence. I exploit various kinds of evidence to provisionally assess the scale of the problem and argue that standard discussions of the reproducibility crisis in psychology and neuroscience ought to pay more attention to its communication dimension.

The paper unfolds in the following manner. In section 1, I introduce the traditional distinction between the notions of replicability and reproducibility in experimental research. Based on a methodological analysis of the requirements of replicability and reproducibility, I show that the latter requirement, unlike the former, is inextricably linked with the norms of scientific reporting, especially the norm of descriptive completeness. In section 2, I describe the standard methods of estimating reproducibility of experimental research and show how they are applied to experimental psychology and neuroscience. I conclude that these methods ignore or even minimize the communication dimension of reproducibility, which suggests a possible blind spot in the existing assessments. I close the section by hinting at the first source of indirect evidence regarding the actual impact of communication failures on reproducibility – namely, the literature on the replicability and reproducibility crisis in computational modeling. As it will transpire, both kinds of crisis in computational modeling have been caused by a widespread failure to adhere to the norm of descriptive completeness. Thus, section 3 introduces the notions of model repeatability, model replicability and model reproducibility, and describes the methodological functions of model replications and model reproductions. It also characterizes the standard methods used to estimate the magnitude of the crisis and, based on available studies, provides the estimates for computational neuroscience. The crisis, it turns out, is serious. In section 4, I return to experimental psychology and neuroscience. Having shown that inadequate reporting has caused a replicability and reproducibility crisis in computational modeling, I look for empirical evidence regarding the claim that something similar may be happening in experimental research. I give two indirect arguments appealing to empirically confirmed deficiencies in the reporting practices of psychologists. It turns out that researchers are less than eager to share experimental data and they also display a remarkable tendency to misreport statistical information. These attitudes, though related to communication, need not impact research reproducibility, however. Therefore, in section 5, I bring up some direct evidence for my

central claim – I discuss four empirical studies: (a) by Grant et al. (2013), who investigate quality of reporting in psychological and social intervention trials in clinical psychology, criminology, education and social work, (b) by Robb et al. (2018), who examine compliance with field specific reporting guidelines in the latest studies of music-based interventions in healthcare, (c) by Brown et al. (2014), who investigate descriptive completeness of psychological papers published in 2010, and (d) by Carp (2012), who focuses on reports of fMRI studies. They all indicate that quality of reporting relevant to reproducibility is poor: roughly over 50% of the reports are seriously deficient in this respect and even about 80–90% of the studies may be irreproducible due to incomplete reporting. I close the paper with a brief summary of the argument and a few remarks on the significance of the communication dimension to the reproducibility crisis.

1. Replication Studies and the Replicability vs. Reproducibility Requirements in Experimental Research

Given the aim of this paper, the most important difference we need to draw between various kinds of replication studies concerns authorship. Thus, I define an *independent replication* as one conducted by independent investigators – i.e., a research team that does not overlap with the authors of the original study. By contrast, an *internal replication* of experimental research is performed by a research team that does overlap with the authors of the original study. Both independent and internal replications are said to be *successful* when the results they yield are sufficiently similar to those of the original study; what precisely counts as sufficiently similar varies with the field. I say that research is *replicable* to the extent that its successive internal replication attempts are likely to be successful. By analogy, I say that experimental research is *reproducible* to the extent that its independent replication attempts are likely to be successful. It follows that there can be two corresponding requirements imposed on experimental research: first, that it be replicable and, second, that it be reproducible.

The other significant distinction regarding repeat experiments concerns the extent to which a replication attempts to reflect the original study. I use the term *direct replication* for studies intended to recreate as closely as possible all the parameters of the original study. Because it is often impossible to duplicate a psychology or neuroscience study down to the smallest detail, any amendment to the original procedure should be justifiable by appeal to the best contemporary understanding of the phenomena under investigation. I use the term *conceptual replication* for replication studies intended to extend the original result or make it more robust by modifying some parameters of the study under investigation. Conceptual replications are not the focus of this paper, so whenever I talk about replications without qualifying them as conceptual, I am referring to direct replications.

Although rarely mentioned in the literature before the 11th century AD, the requirement of replicability may be as old as the experimental method itself (Steinle 2016). A practitioner of empirical science quickly learns how many things can, and often do, go wrong during an experiment. Therefore, it is plausible to suppose that early scientists had developed the habit of repeating their experiments on numerous occasions in order to convince themselves that the outcomes they obtained were sufficiently consistent. Unsurprisingly, assertions of replicability were also used as a rhetorical device, intended to persuade the readers of a scientific text that the claims made by the author were true. For example, the Arab scholar Ibn-al-Haytham, known in Europe as Alhazen, often concluded his experiment descriptions with the words “this is always found to be so, with no variation or change” and Galileo Galilei stressed that he had performed his inclined plane experiments “a full hundred times” (see Steinle 2016, p. 43).

The requirement of reproducibility is much more recent by comparison. It was first invoked as an essential element of the experimental method both in *Saggi di naturali esperienze* (1667), published by the Accademia del Cimento in Florence, and in Thomas Sprat's *History of the Royal Society of London for the Improving of Natural Knowledge* (1667) (cf. Steinle 2016, pp. 46–47). The link between the requirement of reproducibility and the appearance of the first institutions of modern science was no historical accident. The idea of reproducibility was part and parcel of a system of social inventions that had transformed the investigation of nature in the 17th century into a massively collaborative endeavor that has since yielded unprecedented rewards.

As already mentioned, the idea that a credible experimental finding must be replicable came with the awareness that an investigator has to control a wide range of experimental parameters in order to obtain consistent outcomes. Assuming that the researcher is not sloppy, irreplicability either signals indeterminacy in the world or reveals that the researcher has an insufficient understanding of the phenomena under investigation. In either case, the primary advantage of replicating research is that it guards the investigator against accepting certain kinds of false positives and guides further research. From an investigator's point of view, a result that replicates is stable enough to warrant an explanation. By contrast, a result that does not replicate indicates that more research is needed to determine the source of the problem.

The requirement of replicability serves the individual investigator rather than the research community as a whole. The policy of attempting to replicate obtained findings allows the scientist to reject empirically unstable results, but does not necessarily shape the way she communicates her research to others. Although keeping a record of how an experiment was conducted facilitates replicability, it is not its necessary condition. The investigator may be able to replicate a finding by remembering a few distinguishing features of the experimental setup and relying on habit to recreate it. More to the point, she may produce highly replicable research and be secretive about how she is able to do so. In fact, following the requirement of replicability and keeping the necessary know-how hidden from others was typical of much of early experimental practice.

The requirement of reproducibility, by contrast, presupposes communication. It would simply be impossible for an independent investigator to recreate a study unless its author provided a complete and accurate description of the experimental procedure. The fact that a finding can only be reproduced if appropriately communicated to an independent investigator entails that the requirement of reproducibility imposes some general constraints on scientific reporting. Not only does it reinforce the traditional qualities of effective communication, such as accuracy, clarity, precision and conciseness, but it also demands that the research report contain all the information necessary for a competent reader to be able to recreate the study.

This link between the requirement of reproducibility and the norms of scientific reporting, including the norm of descriptive completeness, is immensely important. Ensuring reproducibility does not only help to reduce the rate of false positives in the literature but it also facilitates accumulation of knowledge and boosts scientific progress by making research fully available to anyone who wishes to build on it. Finally, reproducibility is necessary for developing technologies based on scientific findings. This is so even if we lack theoretical understanding of the phenomena under investigation, as in the case of high-temperature superconductivity research, which has led to a number of valuable inventions although the nature of the effect remains unexplained (see Steinle 2016, p. 50).

In light of these remarks, it is clear that reproducibility has two separate but interdependent aspects. Under the epistemic aspect, reproducibility is seen to involve an experimental stability that lends credibility to specific findings and, derivatively, to a whole field of research. Experimental stability is a good proxy for empirical truth although, strictly speaking, the two are logically independent. First, a field of study may produce highly stable results that are in fact false – say, due to a widespread use of a statistical analysis package containing an undiscovered bug. Second, it is possible that most findings generated within a field of research are true but irreproducible (this may be the kind of scenario envisaged by those who are unmoved by the reproducibility crisis). Under the communication aspect, reproducibility involves the publishing of experiment descriptions that meet the norms of scientific reporting. Violation of any of these norms affects the ability of independent investigators to assess or reproduce a finding. The norm of descriptive completeness is crucial in this latter connection because, if it is not being followed, researchers are unable to find information necessary to duplicate the study. As a result, even a perfectly stable experimental finding cannot be validated or used to develop new technologies.

Needless to say, other norms of scientific communication, besides the norm of descriptive completeness, are also relevant to reproducibility. These include the norm of accuracy and the norm of clarity. Reports that either misrepresent the experimental procedure or are plagued by ambiguity can make it impossible for an independent investigator to recreate the study. What is perhaps less obvious, providing too much information – i.e., information that is not necessary to duplicate a study – can have similar effects (Miłkowski, Hohol, & Hensel 2018). Although, in what follows, I touch on whether these norms are being followed, I focus on the norm of descriptive completeness because its violations seem to have the largest impact on the current reproducibility crisis.

2. Assessments of Reproducibility in Experimental Psychology and Neuroscience – the Epistemic Dimension

Like any empirical feature, reproducibility can be measured directly or indirectly. The direct method of assessing reproducibility is simple. It consists in actually trying to independently replicate research. The trouble is that, in many fields, replication attempts are too rare to warrant generalization. According to Makel, Plucker & Hegarty (2012), who searched publication data from the hundred psychological journals with the highest five-year impact factor, the string "replicat*" appears in 1.57% of papers published in the years 1900–2009. Based on a closer analysis of a random sample, Makel and colleagues estimated that only about 64.8% of the papers mentioning the word "replication" are actual replications. This puts the overall rate of replication attempts at 1.07% (though, admittedly, in the period after 2000, it is estimated at 1.96%). However, out of those 1.07%, only 47.1% were performed by independent investigators and 14% were direct rather than conceptual replications. When coupled with the publication bias, a large number of underpowered conceptual replication studies can, in fact, create an illusion of robustness (Romero 2019). Given these figures, it is nearly impossible to arrive at a credible estimate of reproducibility based on prior replication attempts. Indeed, although Makel, Plucker, & Hegarty (2012) found that the reproducibility rate in psychology between 1900 and 2009 was 64.6% (69.4% in 2000–2009), their estimate is probably off the mark as it does not distinguish between direct and conceptual replications and relies on a biased literature.

Another method for directly assessing the reproducibility rate of a discipline involves selecting a representative sample of studies and attempting to replicate them. The largest such effort in behavioral sciences is the Reproducibility Project: Psychology conducted by the Open Science Collaboration (2015). Its results have been

widely publicized: only 36-47% randomly selected recent studies from three of the most prestigious psychological journals have been successfully replicated (n = 100). The estimated reproducibility rate differs according to the field: e.g., it is about 50% in cognitive psychology and about 25% in social psychology. The average effect size was 50% smaller than that reported in the original study.

We may also glean important information from studies investigating smaller numbers of effects, such as the Many Labs 2 replication project, which examined 28 psychological effects reported in both classic and contemporary studies. Its main objective was to assess variation in reproducibility across samples and settings rather than estimate the reproducibility rate, so the sample was not random: some of the effects under scrutiny had been known to be highly reproducible and the selection criteria included simplicity of design and brevity of experimental procedure, which arguably favored easily reproducible results. The outcome: 50% of the effects have replicated ($p < .0001$) (Klein el al. 2018). A related effort, with similar selection criteria, has replicated 3 out 10 effects (Ebersole et al. 2016). Both replication projects also showed considerable effect size inflation in original reports.

Large-scale replication projects are logistically and financially demanding, even if the studies under investigation are relatively cheap and simple. The challenges are even greater in neuroscience where research is more resource-intensive. Boekel et al. (2015) conducted a preregistered multi-study replication attempt of 17 structural brain-behavior (SBB) associations, such as that individual differences in the number of Facebook friends and real social-network size are positively correlated with grey matter volume in several brain areas. Only one of the correlations (about 6%) has replicated. The sample was only 36 subjects but the conclusion is consistent with that of Kharabian Masouleh et al. (2019), who reanalyzed a large dataset containing brain scans and selected psychological tests.

Given the heterogeneity of both experimental psychology and neuroscience and the difficulties associated with conducting large-scale replication projects, most assessments of the reproducibility crisis are indirect. They are based on estimating the value of certain parameters that are either known or supposed to indicate reproducibility.

A large number of these meta-scientific studies focus on the truth of scientific results. They involve sifting through the published record and applying a wide range of statistical methods to estimate the value of such parameters as the average statistical power and effect size (e.g., Ioannidis 2005; Button et al. 2013; David et al. 2013; Szucs & Ioannidis 2017, Stanley, Carter, & Doucouliagos 2018), the prevalence of various biases (e.g., Fanelli 2009, 2010, Scheel 2019), the flexibility of current research methods (e.g., Patel, Burford, & Ioannidis 2015), etc.

A second group of meta-scientific studies of reproducibility attempt to discover the frequency of behaviors that do not conform to the principles of scientific investigation. These studies are relatively rare and often involve conducting anonymous surveys among researchers, asking them about their conduct, experiences or attitudes (e.g., Anderson, Martinson, & DeVries 2007; Fanelli 2009; Baker 2016; Mobley et al. 2013; Héroux et al. 2017; John, Loewenstein, & Prelec 2017). Although there are surveys that ask scientists about selective reporting, which obviously violates the norm of descriptive completeness, they do not generally target the communication dimension of reproducibility.

What is the reproducibility rate in experimental psychology and neuroscience? It should come as no surprise that opinions on this point vary. Despite our best efforts, we still lack sufficient data to assess the situation, and the data we do have are interpreted in a variety of ways because every analysis relies on some controversial assumptions (see Amrhein, Trafimow, & Greenland 2019). Let me illustrate this with a vivid example. On one

side of the spectrum, Gilbert et al. (2016) argue that the empirical evidence obtained by the Open Science Collaboration (2015) is actually consistent with the reproducibility rate in psychology being around 85%. This is because, they claim, the replicability benchmarks adopted by the Open Science Collaboration did not take into account the infidelities of the replications (some of which may even suggest bias) and the replications were underpowered because power calculations were based on considerably inflated effect sizes reported by original investigators (cf. Anderson et al. 2016 for a rejoinder; for more on the question of power in large-scale replication efforts, see Camerer et al. 2018). On the other extreme, Johnson et al. (2017), who performed a reanalysis of the Open Science Collaboration (2015) dataset, claim that the pre-study odds of experimental hypotheses in psychology – i.e., the probability that a hypothesis being tested is actually false – are likely below 10% and the rate of false positives is likely over 90% (cf. Szucs & Ioannidis 2017 for similar estimates based on a different methodology and Bird 2018 for a general argument that low reproducibility may be the result of low pre-study odds even in the absence of questionable research practices).

It is not my aim here to investigate which proposed estimates are the most plausible and why. Instead, I want to stress that the methods of assessment targeting what I call the epistemic dimension of the reproducibility crisis ignore the communication dimension. In fact, multi-study replication projects are typically designed so as to minimize the impact of inadequate reporting on estimated reproducibility. Replication teams in the Open Science Collaboration contacted the original authors to obtain original study materials, request "any important information about the methodology that may not have appeared in the original article" and "share the methods draft for comments and suggested edits" (Reproducibility Project: Psychology Researcher Guide, https://osf.io/ru689/). Many Labs 2 and 3 had a similar protocol (Klein et al. 2018, p. 448; Ebersole et al. 2016, p. 70). Boekel et al. (2015) also solicited materials from original authors. Interestingly, their initial plan had been to replicate 9 studies. They ended up with only 5 because, among other things, the original authors of 3 studies either failed to supply ROI masks or the masks they sent did not match the coordinates reported in the papers.

Since the methods of assessment described so far cannot detect the possible impact of descriptive incompleteness on reproducibility, it is at least imaginable that this has led the community to consistently underestimate the magnitude of the reproducibility crisis in experimental psychology and neuroscience. I begin to explore this possibility in the next section, where I discuss the prevalence of incomplete reporting in computational modeling. The data suggest that computational sciences are experiencing a crisis similar to that in experimental psychology and neuroscience. However, unlike in experimental research, there, the main culprit is believed to be inadequate reporting practices. This constitutes the first piece of indirect empirical evidence I usher in to support the hypothesis that communication problems are also affecting the reproducibility of experimental psychology and neuroscience.

3. The Communication Dimension of the Reproducibility Crisis: The Case of Computational Modeling

There are three senses in which computational models can be duplicated. First, a model is said to be *repeatable* to the extent that its author can rerun it herself at a later date and obtain the same output given the same input data. Second, a model is said to be *replicable* to the extent that an independent researcher can run the original source code on original input data and obtain the same output as the original investigator. Third, a model is said to be

reproducible to the extent that an independent researcher can reimplement it without using the original source code and, given the same input, obtain the same output as the original investigator (Delling et al. 2016).

Note that the contrast between replicability and reproducibility of experimental research, drawn in terms of authorship, breaks down when applied to computational modeling: both model replication and model reproduction are performed by an independent investigator. This is unfortunate but inevitable if we want to keep to standard usage. Speaking of terminology, I should also warn the reader that there are two competing ways of using the terms "model replicability" and "model reproducibility" in the literature. I have chosen to follow a convention inspired by Drummond (2009) and endorsed by the Association for Computing Machinery. According to the other tradition, the meanings of the two terms are reversed: i.e., what I defined as model replicability is referred to using the term "model reproducibility" and vice versa (cf. Plesser 2018; Miłkowski, Hensel, & Hohol 2018). Needless to say, there are also writers who use the terms "model replicability" and "model reproducibility" more or less interchangeably (see Barba 2018).

Someone unfamiliar with computational modeling may be surprised to learn that neither repeatability nor replicability is easy to accomplish. After all, computers are deterministic machines, aren't they? In fact, however, the sheer complexity of the models and the rapid changes of programming environments make rerunning a computer simulation and actually getting the same output a potentially tedious exercise, especially after a year or two. In order to ensure repeatability, the modeler has to document every step of the computation leading from input data to the final result because the slightest variation may affect the output. This includes keeping track of software versions as well as the exact versions of scripts and parameters used. Because human memory is limited, it is also important to store descriptions connecting successive versions of the code to the conceptual model the code is intended to implement (Sandve et al. 2013).

Making sure that computational research is repeatable and replicable is as much about record keeping as it is about quality of the research. And, just as in experimental studies, keeping detailed notes facilitates error correction (Donoho et al. 2009). In a sense, researchers who want to build repeatable models need to be able to communicate effectively with the future versions of themselves, and researchers who want to produce replicable models need to communicate effectively with other modelers. Insofar as a future version of the original investigator is similar to other modelers, replicability can be identified with repeatability plus public availability of the code and data. This is why nine out of the ten rules for replicable computational research recommended by Sandve et al. (2013) focus on repeatability. Note, however, that the idea that notes should be written as if they were addressed to someone other than the author – as if their intended reader were a different modeler – had been shaped by the requirement of reproducibility.

There are two dimensions along which computational models are evaluated. The first concerns what is known as *correctness*, which is a relation between an implementation and the conceptual model the implementation is taken to embody. Checking for correctness is called *model verification*. Verifying a computational model is basically demonstrating that the code operates as intended or, as Balci put it, "model verification deals with building the model *right*" (1997, p. 135). The second dimension is *validity*, also called *faithfulness* (and sometimes *fidelity*), which is a relation between the model and the system it represents. To *validate* a model is to demonstrate that it describes the behavior of the source system (the target phenomenon) at the level required by the objectives of the study (Zeigler, Muzy, & Kofman 2019, pp. 32–33). As Balci put it, "validation deals with building the *right* model"

(1997, p. 135). Verification and validation are conducted throughout the model development process but, since no method of model analysis guarantees either perfect verification or perfect validation, both activities extend beyond that.

Replicating a model provides a method of its verification (Rand & Wilensky 2006) and, although there are other methods as well, ensuring model replicability has some important advantages besides offering an opportunity to discover previously undetected errors. Replicability enables researchers to reuse and modify the original code and also contributes to model validation by making it possible to run the simulation on new datasets. Both of these contribute to theoretical progress.

Performing a reproduction attempt can also confirm or disconfirm the computational model's correctness. A failed attempt at reconstructing a model based on its verbal description indicates that either the description is flawed (inaccurate or incomplete) or the original implementation was incorrect (Cooper & Guest 2014). Yet model reproductions also serve a different, more important function. They allow us to separate the theoretical assumptions of the conceptual model from the extrinsic properties of its implementation – i.e., features of the implemented model that are not intended as part of the theoretical description of the world provided by the conceptual model. The two kinds of properties often get conflated because most researchers do not specify the conceptual model independently of its implementation (Cooper & Guest 2014; McClelland 2009). Thus, reproductions improve our understanding of the model and contribute to its validation, which is more theoretically significant than model verification (Miłkowski, Hensel, & Hohol 2018; Drummond 2009).

Given the importance of model replicability and reproducibility, it is unsettling that, like experimental psychology and experimental neuroscience, computational modeling is experiencing a confidence crisis associated with poor replicability and reproducibility of published research. And, same as with experimental science, there are two kinds of methods that can be used to assess the rates of replicability and reproducibility. The direct methods of assessment involve performing model replications or model reproductions whereas the indirect methods focus on estimating the value of various parameters known or supposed to affect model replicability or model reproducibility.

The direct methods of assessment are even more rarely used in computational modeling than in experimental science. There is the journal *ReScience C*, established in 2015, dedicated to publishing model reproduction studies (Rougier et al. 2017) but there are no initiatives like the Open Science Collaboration that aim to estimate the rates of replicability or reproducibility of computational models across the discipline. Although some research teams make consistent attempts to reproduce models in their field of interest, the results of their efforts do not yet add up to anything resembling a general picture of the crisis. For example, Manninen et al. (2018) have tried to reimplement two models of neural signal transduction, two spiking neuronal network models and eight models of astrocyte activity. They managed to obtain all the original results for four studies (25%). As for model replicability, its direct assessments are almost never made, which is not a serious problem, however, because indirect assessments have much more scope and provide a good approximation of the upper bound on estimates of replicability. The study by Manninen et al. (2018) is the only attempt I know of to directly assess model replicability. The team evaluated the replicability of 10 spiking neuronal network models. Only 3 of the models were available on line – the proportion of replicated results was over 60% for all three studies.

Most information we have about the replicability crisis in computational studies comes from estimates of the prevalence of its necessary condition: public availability of the code and data. According to a survey by Gundersen, only 6% out of the 400 papers delivered at two top AI conferences organized in recent years have made the source code for the algorithms publicly available (Hutson 2018). As to the field of computational simulation, Stodden et al. (2018) have recently evaluated the effectiveness of a share-artefacts-on-demand policy introduced by the journal *Science* in 2011. It turned out that, out of the 204 computational studies selected by Stodden and colleagues, 24 (about 12%) provided the source code and data by means of an external link or supplementary materials. After contacting the authors of the remaining studies, Stodden and colleagues were able to obtain artifacts for another 24% of the papers. According to a survey conducted by Miłkowski, Hensel and Hohol (2018), the situation in computational neuroscience is slightly better, though far from ideal: 32% out of the 242 papers published between January 1, 2016, and September 26, 2018, in three prominent computational neuroscience journals provided the source code for the model. This matches almost exactly the proportion of available artefacts reported by Manninen et al. (2018) for the ten spiking neuronal network models they attempted to replicate.

Unfortunately, we do not have similar data on the reproducibility crisis, which is both understudied and vastly underappreciated by the modeling community. I venture that model reproducibility rates are lower than those of model replicability because, other things being equal, model reproductions are harder and more time-consuming to perform than are model replications. Information necessary to reconstruct a model based on its verbal description is interspersed throughout the paper, some of it is ambiguous or missing, and the equations and figures contain mistakes which a researcher attempting a reimplementation must try to correct. The time it took Manninen et al. (2018) to reproduce a model ranged between 2 hours and over 2 weeks – by contrast, a replication took them between no time at all and 2 days. We can also predict that, without a radical change in the way computational modeling is being done and evaluated, both crises are going to get worse because replications and reproductions become increasingly difficult and time-consuming as the models get larger and more complex – which of course they are.

Despite paucity of data on the extent of the reproducibility crisis in computational modeling, the community is becoming increasingly aware of the need to improve descriptive completeness of reporting. The first step has already been taken. Various subfields of computational modeling have developed and endorsed reporting standards. For example, Nordlie, Gewaltig and Plesser (2009) provided detailed guidelines and recommendations regarding descriptions of neural network models. This is an important development because it shows that researchers in some areas have noticed how difficult it may be to decide what to include in a research report and how exactly to go about it, especially in a rapidly growing field. Which information about a neural network model ought to be included in the main body of the paper and which can be relegated to supplementary materials? What if there is a space limit? How to construct complete and unambiguous diagrams and tables? In any subfield, there are a number of such seemingly mundane questions that must be answered in an informed and reasoned way if research is to be effectively communicated and thereby potentially reproducible (see Miłkowski, Hensel, Hohol 2018).

It is still an open question whether standards of reporting developed by some communities in computational modeling are actually being met. None of the papers describing the ten spiking neuronal network models

investigated by Manninen and colleagues presented model parameters in the format proposed by Nordlie, Gewaltig and Plesser (2009) (Manninen et al. 2018, p. 12).

Having described the nature of the replicability and reproducibility crisis in computational modeling and assessed its scale based on available data, let me return to experimental psychology and neuroscience. The second indirect empirical argument for my hypothesis that the two disciplines are suffering from the reproducibility crisis in large part because of the researchers' failure to adhere to the norm of descriptive completeness appeals to evidence of two worrying patterns of behavior concerning communication.

4. Two Worrying Patterns of Behavior in Experimental Psychology and Neuroscience

There are at least two attitudes that involve withholding information from fellow researchers or misrepresenting it but do not directly affect reproducibility. The prevalence of these attitudes affects experimental research by giving rise to something akin to the replicability crisis in computational modeling. In computational modeling, failure to share code and data limits the ability of independent investigators to assess the correctness of a model's implementation and inhibits scientific progress by making it impossible to reuse code. In experimental research, failing to share experimental data limits the ability of independent investigators to assess the validity of the arguments and inferences made by the original authors and, at least in the case of neuroimaging data, inhibits progress by preventing reuse. Note also that the methodological differences between model replication and model reproduction carry over to experimental research. When reanalyzing the data of an empirical study, one can run the original analysis script and check how well the outputs match with reported outcomes (analytic/computational replication), or write a new script based on the description provided in the original paper and compare its outputs with reported outcomes (analytic/computational reproduction).

Kidwell et al. (2016), who investigated articles published in 2012 and 2013 by five prominent psychology journals, found that less than 3% of the papers declared public availability of the data (and most of the datasets turned out to be unusable or incomplete). A similar problem occurs in neuroscience, where less than a few percent of fMRI scans are available in public repositories (Poline et al. 2012). Furthermore, researchers have been known to refuse to share data. In 2005, Wicherts and colleagues contacted the authors of 141 papers published in four APA journals, which require the authors to share data on request. Only 27% of the authors complied (Wicherts et al. 2006).

On the upside, we are seeing signs of improvement. When Vanpaemel et al. (2015) requested data from the authors of almost 400 papers published by APA journals, 38% of the authors complied. Furthermore, data sharing is now in vogue and the consequences of this are beginning to be felt. Nelson, Simmons and Simonshon (2018) have even gone so far as to proclaim a renaissance of psychological science. However, most data are not in yet and the current enthusiasm for open science may prove to be a short-lived phenomenon. What is relevant to my argument is that researchers, including those in experimental psychology and neuroscience, have persistently exhibited a marked reluctance to share data. This reluctance seems to emerge spontaneously in many fields despite declared adherence to Mertonian norms of scientific investigation (Anderson, Martinson, & DeVries 2007; Houtkoop et al. 2017).

A second behavioral pattern associated with communication is the misrepresentation of statistical results. Although it rarely renders replication attempts impossible, it can increase the rate of false positives in the literature and make it difficult or even impossible to assess the credibility and theoretical significance of published findings, which, in the long run, is bound to affect the reproducibility rate in a whole field (cf. Muthukrishna & Henrich 2019). Garcia-

Berthou and Alcaraz (2004) found that over 11% of *Nature* papers using NHST reported incongruent statistical results, meaning that reported *p* values were inconsistent with other statistical information provided in the text. Bakker & Wicherts (2011) performed a similar study focusing on psychology. They found that about 18% of statistical results were misreported and, more importantly, about 15% of the papers contained reporting errors that affected at least one statistical conclusion. Nuijten et al. (2016) further confirmed these findings by performing an automated analysis of 30,717 papers published between 1985 and 2013 by eight top psychology journals: 49.6% of the papers using NHST contained at least one inconsistency and 12.9% contained an inconsistency that affected a statistical conclusion. Given the perceived role of statistical significance in psychological inference, the frequency of this kind of reporting error is striking. It would seem that *p* values should be the subject of especially close scrutiny both by the authors of a paper and its reviewers. Although some of the errors are likely intentional, a great deal of them are probably honest mistakes that have not been caught before publication.

The two behavioral patterns I have discussed – namely, not sharing data and misreporting statistical information – are only indirectly relevant to the communication aspect of the reproducibility crisis in experimental psychology and neuroscience because neither a lack of original data nor many of the mistakes in statistical information prevents direct independent replication. My argument here is that if researchers withhold or misrepresent one kind of information (irrelevant to reproducibility) then they are likely to withhold or misrepresent other kinds of information as well (potentially relevant to reproducibility). For example, out of 326 empirical studies published between 2012 and May, 2015, in four leading psychology journals, almost 20% contained a materials availability statement but of the materials "reportedly available at a website or repository, 60% (N = 27) were actually available, 46.7% (N = 21) were correct materials, 33.3% (N = 15) were usable materials, and 13.3% (N = 6) were complete materials" (Kidwell et al. 2016, p. 9). I would submit that the same, *mutatis mutandis*, applies to reporting infidelities but they are much more difficult to detect.

Needless to say, both inferences are based on folk psychology and do not have sufficient empirical support. However, the corresponding psychological constructs – i.e., the dimensions of conscientiousness and honesty/dishonesty are largely domain-independent and manifest themselves in numerous aspects of life (see, e.g., Roberts et al. 2009; Mazar & Arieli 2015).

5. Direct Evidence for Widespread Failure to Provide Complete Research Descriptions in Experimental Psychology and Neuroscience

It is time to confront our main question head on: "What proportion of published research reports in experimental psychology and neuroscience do not contain all the information a knowledgeable investigator would need in order to recreate the reported study?"

Naturally, a reasonable answer must involve a review of the literature on those aspects of quality of reporting that are likely to influence reproducibility. Studies of this kind, though not necessarily targeting reproducibility, are relatively common in biomedical science. They usually compare a corpus of publications against an established set of guidelines, developed by experts, informed by research and consistently updated. Some guidelines, such as Consolidated Standards of Reporting Trials (CONSORT, http://www.consort-statement.org/), Animal Research: Reporting of *In Vivo* Experiments (ARRIVE, https://www.nc3rs.org.uk/arrive-guidelines) and The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE, https://www.strobe-

statement.org/index.php?id=available-checklists), apply to many fields and study protocols, whereas others, such as Standards for Reporting Interventions in Clinical Trials of Acupuncture (STRICTA, https://www.stricta.info/), are much more specific (for more details see https://www.equator-network.org/, which provides a comprehensive list of reporting standards in health research). The guidelines are presented in the form of checklists, flowcharts and, increasingly, full-text explanations and elaborations (e.g., Moher et al. 2010).

Although there is much room for improvement, both psychology and neuroscience have developed and, to an extent, adopted a number of reporting standards. The Publications and Communications Board of APA appointed the first Working Group on Journal Article Reporting Standards (JARS) in 2006. The group's report was published as an article (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and the guidelines were incorporated, in the form of four tables and a flowchart, into the sixth edition of the *Publication manual of the American Psychological Association* (American Psychological Association, 2010). The recommended standards related to: (1) reports of new data collections regardless of research design, (2) studies with a manipulation or intervention, (3) studies using random and nonrandom assignment of participants to experimental groups and (4) meta-analysis reporting standards (MARS). The guidelines have recently been updated and expanded (Appelbaum et al. 2018). Neuroscientists follow biomedical and psychological reporting guidelines but also have specific standards of their own, such as Picton et al. (2000).

The most important developments cited by the authors of JARS as providing motivation for the development of reporting standards were the movement toward evidence-based practice in medicine and education and an increasing reliance on research syntheses, especially meta-analyses, in behavioral and social sciences. This highlights that reporting standards are typically formulated with more than one aim in mind, which means that not all research into quality of reporting will be relevant to reproducibility. The functions of reporting standards that have little or no impact on reproducibility include improving the readers' ability to accurately interpret the research and facilitating application of datamining techniques. For example, it is largely irrelevant to reproducibility if most papers in a field do not contain some specific information in their titles or abstracts, though it may be relevant to datamining and, derivatively, to meta-analyses.

Research into how well the psychological and neuroscientific literature actually conforms to reporting guidelines is scarce, especially if we restrict our attention to adherence to standards potentially relevant to reproducibility. Alas, the situation is not hopeless. There are a handful of studies that can shed some light on our question. Most of them are found at the intersection of experimental psychology and neuroscience, on the one hand, and computer science and biomedicine, on the other. The sample is far from representative but beggars can't be choosers. Below, I discuss the findings of four such studies.

The first study, by Grant et al. (2013), investigated adherence to a variety of published standards in controlled randomized trials (CRTs) of social and psychological interventions. It covered clinical psychology, criminology, education and social work, targeting reports "representative of the best trial research in these disciplines" (Grant et al. 2013, p. 8). Given its large scope, it can give us a general idea of guideline effectiveness in reports of CRTs of non-pharmacological treatments. The findings did not vary significantly with discipline. Average compliance with all standards across the disciplines was 42%.

Adherence to reporting standards related to internal and external validity was 38% and 47%, respectively. To zoom in on some of the details relevant to reproducibility: while 60% of the studies reported trial eligibility criteria, the majority failed to list all inclusion and exclusion criteria, and although 71% of the studies provided the number of participants assigned to each condition, compliance with reporting guidelines concerning descriptions of random sequence generation, allocation concealment and blinding ranged between 15 and 23%. This suggests that descriptively incomplete reporting has a considerable impact on the reproducibility of CRTs of psychological and social interventions as well over 50% of reports published in high-impact journals are seriously deficient and, apparently, not more than 15% of the reports are complete.

The second study, by Robb et al. (2018), has less scope but provides much more detail. It examines reporting quality of music interventions in healthcare by comparing the latest literature (187 papers from 2010–2015) against a set of guidelines specific to the field (Robb, Burns, & Carpenter 2011). Most of these guidelines were expressly designed to improve reproducibility and translatability to treatment. The guidelines apply to the following aspects of the interventions: theoretical background (rationale for selected stimulus, specification of how the stimulus is expected to affect the targeted outcome), content (details of the intervention: intervention strategies, who selected the music, what kind of music it was, how the stimulus was delivered, etc.), delivery schedule (number, duration and frequency of sessions), interventionist (the number, qualifications and credentials of the persons who delivered the intervention), treatment fidelity (strategies of ensuring that the conditions were delivered as intended), setting (location, privacy level and ambient sound) and unit of delivery (individuals or groups).

As expected, information in some categories was reported by most studies. Thus, 100% of the reports specified the intervention strategies (e.g., music listening, songwriting, rhythmic auditory stimulation, etc.), 91% specified the unit of delivery and 89% – the person selecting the music (e.g., patient or investigator); 81% of the studies provided a complete description of the delivery schedule. However, only 53% of the reports described the theoretical background of the intervention, 34% specified music and non-music materials used, 23% provided information on volume of the music, 20% reported on treatment fidelity, 18% characterized the level of privacy and 12% described the level of ambient sound; 15% of the studies using recorded music provided reference for sheet music or sound recording. Overall, compliance was below 50% in descriptions of four out of the seven aspects covered by the guidelines. As each of these aspects, with the possible exception of theoretical background, is directly relevant to reproducibility, it follows, again, that at least 85% of recent studies published in the field of music intervention may be irreproducible due to incomplete reporting.

Unlike the first two studies I have discussed, the third one, by Brown et al. (2014), assessed the psychological literature in terms of standards that had not been explicitly adopted by the community. Brown and colleagues began by creating a comprehensive list of information needed by a naïve researcher to replicate a psychological study. The list, which they called The Replicability and Meta-Analytic Suitability Inventory (RAMSI), was based on Mill's canons, various extant standards, such as CONSORT, and empirical studies into factors affecting the outcomes of psychological research. With RAMSI in hand, the investigators examined 1,083 quantitative research reports published during 2010 by the five highest, middle and lowest ranking psychology journals as judged using the top 100 ISI Web of Science 5-year impact factor.

RAMSI consists of five groups of items: method, participants, assessor, experimenter and results. Brown et al. (2014) found that quality of reporting was best in the category of results (descriptive completeness was 65–69%)

and worst in experimenters (11–16%). Overall scores of descriptive completeness were low (29–32%) regardless of journal rank. Although, due to the method of its development, RAMSI is open to the charge of arbitrariness, the study seems to offer a unique insight into the reporting practices in psychology relevant to reproducibility.

My fourth and final example comes from neuroscience, where Poldrack et al. (2008) presented a detailed reporting standard of fMRI research, providing both a set of general principles and a checklist of parameters any fMRI study description should contain. Given the noisiness of fMRI measurements, the value of a majority of the numerous study parameters specified by Poldrack et al. (2008) can arguably affect the final outcome of a study.

Recently, Carp (2012) has investigated adherence to this standard. He selected 241 fMRI papers drawn from 68 neuroscientific journals, such as *PLoS ONE*, *NeuroImage* and *Cerebral Cortex*, and compared them against Poldrack's et al. (2008) checklist. The parameters were grouped into four categories: experimental design, data acquisition, pre-processing and modeling. Each category covered from 6 to 16 parameters. As to experimental design, most studies reported the number of subjects (99.6%), the proportion of female subjects (over 86%), the number of recording sessions (over 98%) and the design type (over 83%). By contrast, only about 6% of the studies reported whether and how the task design was optimized for efficiency, less than 22% of the papers described criteria for excessive head movement and about 28% of the studies reported the number of subjects scanned but excluded from analysis. The list goes on. As to data acquisition, about 42% of the studies reported the voxel dimensions, about 36% the coverage achieved, and just over 22% the order of slice acquisition. Pre-processing was the most poorly described category: only three parameters were specified by a majority of the studies – about 14% of the studies reported the reference slice used for slice-timing correction and only about 1% described the interpolation method used during co-registration. Again, the list goes on. As Carp concludes (2012, p. 297):

> The widespread omission of these parameters from research reports, documented here, poses a serious challenge to researchers who seek to replicate and build on published studies. Changing even a single critical methodological decision may qualitatively alter the results of an experiment; changing many decisions at once may exert profound and unpredictable effects on research outcomes.

6. Concluding Remarks

If my analysis is correct then the reproducibility crisis experienced by experimental psychology and neuroscience stems from two sources. First, the researchers are violating the norms of scientific investigation, which leads to an increase in false positives, which in turn leads to failed replication attempts. We know that this is happening from the standard account. Second, however, the researchers are also violating the norms of scientific reporting that must be followed if research is to be reproducible.

My argumentation for this latter claim proceeded in several stages. I started off by showing the existence of an almost conceptual link between effective communication and reproducibility: namely, that an independent investigator cannot recreate a study without knowing how it was performed. I then described the standard, epistemically oriented methods of assessing the reproducibility of experimental research in psychology and neuroscience, arguing that they neglect the possible contribution of inadequate reporting. Next, I showed how widespread failure to follow the norm of descriptive completeness had in fact led to a replicability and reproducibility crisis in computational modeling. If I were to make an educated guess I would put model replicability at about 20–60% and model reproducibility at about 10–40%. But the point is that the existence of a

replicability and reproducibility crisis in computational modeling demonstrates how much of an impact incomplete reporting can actually have on reproducibility. Having established that, I returned to experimental psychology and neuroscience. Here, I pointed to two worrying behavioral patterns that constitute indirect evidence for the researchers' tendency to produce descriptively incomplete or misleading research reports. And, finally, I discussed four empirical studies into quality of reporting whose results are directly relevant to my claim – they all indicate that failures to properly communicate research affect reproducibility in a major way.

It follows that a good approximation of the scale of the reproducibility crisis in experimental science should involve taking the rate of false positives and adding to it the rate of inadequate research descriptions in the literature. To paraphrase and extend Ioannidis' (2005) dictum: most scientific findings are false but many of the true ones may actually be irreproducible anyway.

Of course, the operation of adding the relative contributions of false positives, on the one hand, and inadequate research descriptions, on the other, remains somewhat metaphorical because we have no idea whether the two dimensions are independent. It is logically possible that most inadequate research reports happen to describe false positive findings, which would make a relatively small fraction of true positives irreproducible. And it is also logically possible that that the opposite is true: that most inadequate research reports happen to describe true positive findings, in which case a relatively large fraction of true positives would be irreproducible. In the absence of data, it is probably reasonable to treat failure to follow the norms of scientific investigation and failure to adhere to the norms of scientific reporting as if they were independent. Be that as it may, the impact of incomplete reporting on research reproducibility in experimental psychology and neuroscience, as well as in other disciplines, seems considerable although much more work needs to be done to properly assess its extent and causes.

My claim that inadequate, and especially incomplete, reporting may actually be a major cause of the current reproducibility crisis in experimental psychology and neuroscience – and that this contribution is relatively neglected – should not be confused with the ideas of the open science movement (see Crüwell et al. 2019 for a review). Although I whole-heartedly support many open science initiatives, my focus in this paper is purely epistemic: I simply want to know what proportion of findings in experimental psychology and neuroscience are likely irreproducible. The open science movement, by contrast, is primarily interested in improving the quality of scientific research and communication in many respects, including but not limited to reproducibility.

Of course, the recommendations are based on evidence, which means that the open science movement is both sensitive to relevant empirical findings and stimulates further meta-scientific research. However, studies into adherence to reporting standards relevant to reproducibility have come mostly from the biomedical sciences, where the communication aspect of the reproducibility crisis seems to have received much more recognition than in experimental psychology and neuroscience. It is no coincidence, for example, that the following quote comes from a paper written mostly by biomedical scientists: "Poor transparency can have very real costs. For example, the Reproducibility Project in Cancer Biology – a major effort to replicate 50 high-impact cancer biology papers – recently had to abandon 32 replication attempts partly because pertinent methodological information about the original studies was not available" (Hardwicke et al. 2020, p. 7). In a sense, then, I am simply applying some insights from computational modeling and biomedicine to experimental psychology and neuroscience. My fear is that, if the reproducibility project mentioned by Hardwicke and colleagues were to be reported, after completion,

from a purely epistemic perspective, the abstract would read something like: *we conducted a replication of 18 high-impact cancer biology studies*.

In fact, Boekel et al. (2015) did write in the abstract: "Here, we attempt to replicate five SBB correlation studies comprising a total of 17 effects" – which is also how I reported the study in section 1. Of course, this does not misrepresent Boekel and colleagues' findings, especially given that the reader learns about the excluded studies from the *Materials and Methods* section of their paper (and about the number of excluded effects from a preregistration protocol, http://confrepneurosci.blogspot.com/2012/06/advanced-methods-and-analyses_26.html). But, I submit, it would not be entirely inadvisable to consider a slightly different formulation as well: "Here, we attempt to replicate eight SBB correlation studies comprising a total of 23 effects".

References

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73*(Sup. 1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839–851. http://dx.doi.org/10.1037/0003-066X.63.9.839

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., … & Della Penna, N. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1037.

Anderson, M. S., Martinson, B. C, & DeVries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics, 2*(4), 3–15. https://doi.org/10.1525/jer.2007.2.4.3

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*, 3–25. doi:10.1037/amp0000191

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

Bakker, M., & Wicherts, J. M. (2011). The misreporting of statistical results in psychology. *Behavior Research Methods, 43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Balci, O. (1997). Verification, validation and accreditation of simulation models. In *Proceedings of the 29th Conference on Winter Simulation*. Atlanta, GA, 135–141.

Barba, L. A. (2018). Terminologies for reproducible research. arXiv preprint arXiv:1802.03311

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex, 66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy051.

Brown, S. D, Farrow, D., Hill, D. F., Gable, J. C, Porter, L. P., & Jakobs, W. J. (2014). A duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science, 9*(6), 626–640. https://doi.org/10.1177/1745691614551749

Button, K. S, Ioannidis, J. P. A., Mokrysz, C., Nosek B. A., Flint, J., Robinson, E. S. J., & Munafó, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(451), 365–376. https://doi.org/10.1038/nrn3475

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*, 637–644, https://doi.org/10.1038/s41562-018-0399-z

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 64*, 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research, 27*, 42–49. https://doi.org/10.1016/j.cogsys.2013.05.001

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., … & Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science: An annotated reading list. *Zeitschrift für Psychologie, 227*, 237–248. https://doi.org/10.1027/2151-2604/a000387

Delling, D., Demetrescu, C., Johnson, D. S., & Vitek, J. (2016). *Rethinking experimental methods in computing*. Schloss Dagstuhl—Leibniz–Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany. https://doi.org/10.4230/DagRep.6.3.24.

Donoho, D. L., Maleki, D., Rahman, I. U., Shahram, M., & Stodden, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science & Engineering, 11*(1), 8–18. https://doi.org/10.1109/MCSE.2009.15

Drummond, D. C. (2009). Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*. Montreal, Canada: National Research Council. http://cogprints.org/7691/. Accessed 17 Feb 2018.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Fanelli, D. (2009). How many scientists fabricate or falsify research: A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), e5738.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology, 4*, 13. https://doi.org/10.1186/1471-2288-4-13

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1036.

Grant, S. P., Mayo-Wilson, E., Melendez-Torres, G. J., & Montgomery, P. (2013). Reporting quality of social and psychological intervention trials: A systematic review of reporting guidelines and trial publications. *PLoS ONE, 8*(5), e65442. https://doi.org/10.1371/journal.pone.00654

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *Royal Society Open Science, 7*, 190806. http://dx.doi.org/10.1098/rsos.190806

Héroux, M. E., Loo, C. K., Taylor, J. L., & Gandevia, S. C. (2017). Questionable science and reproducibility in electrical brain stimulation research. *PLoS ONE, 12*(4), e0175635. https://doi.org/10.1371/journal.pone.0175635

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Hutson, M. (2018). Missing data hinder replication of artificial intelligence studies. *Science*. https://doi.org/10.1126/science.aat3298.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112*(517), 1–10. https://doi.org/10.1080/01621459.2016.1240079

Kharabian Masouleh, S., Eickhoff, S. B., Hoffstaedter, F., Genon, S.; Alzheimer's Disease Neuroimaging Initiative (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *Elife, 8*, e43464, https://doi.org/10.7554/eLife.43464

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency, *PLoS Biology, 14*(5), e1002456.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Manninen, T., Aćimović, J., Havela, R., Teppola, H., & Linne, M.-L. (2018). Challenges in reproducibility, replicability, and comparability of computational models and tools for neuronal and glial networks: Cells and subcellular structures. *Frontiers in Neuroinformatics, 12*, A20. https://doi.org/10.3389/fninf.2018.00020

Mazar, N., & Arieli, D. (2015). Dishonesty in scientific research. *The Journal of Clinical Investigation, 125*(11), 3993–3996. https://doi.org/10.1172/JCI84722

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*(1), 11–38. https://doi.org/10.1111/j.1756-8765.2008.01003.x

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience, 45*(3), 163–172. https://doi.org/10.1007/s10827-018-0702-z

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., & Zwelling, L. (2013). A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS ONE, 8*(5), e63221. https://doi.org/10.1371/journal.pone.0063221

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*, 221–229 (2019). https://doi.org/10.1038/s41562-018-0522-1

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5, 85–90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*, 21–29.

Nordlie, E., Gewaltig, M.-O., & Plesser, H. E. (2009). Towards reproducible descriptions of neuronal network models. *PLoS Computational Biology, 5*(8), e1000456. https://doi.org/10.1371/journal.pcbi.1000456.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors psychology. *Behavior Research Methods, 48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology, 68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., ... , & Taylor, M. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37*(2), 127-152. https://doi.org/10.1111/1469-8986.3720127

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics, 11*, 76. https://doi.org/10.3389/fninf.2017.00076.

Poline, J.-B., Breeze, J., Ghosh, S., Gorgolewski, K., Halchenko, Y., Hanke, M., Helmer, K., Marcus D., Poldrack, R., Schwartz, Y., Ashburner, J., & Kennedy, D. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics, 6*, Art. 9. https://doi.org/10.3389/fninf.2012.00009

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage, 40*(2), 409–414. https://doi.org/10.1016/j.neuroimage.2007.11.048

Rand, W., & Wilensky, U. (2006). Verification and validation through replication: A case study using Axelrod and Hammond's ethnocentrism model. *North American Association for Computational Social and Organization Sciences (NAACSOS)*, 1–6.

Robb, S. L., Burns, D. S., & Carpenter, J. S. (2011). Reporting guidelines for music-based interventions. *Journal of Health Psychology, 16*(2), 342–352. http://dx.doi.org/10.1177/

Robb, S. L., Hanson-Abromeit, D., May, L., Hernandez-Ruiz, E., Allison, M., Beloat, A., Daugherty, S., Kurtz, R., Ott, A., Oladimeji Oyedele, O., Polasik, S., Rager, A., Rifkin, J., & Wolf, E. (2018). Reporting quality of music intervention research in healthcare: A systematic review. *Complementary Therapies in Medicine, 38*, 24–41. https://doi.org/10.1016/j.ctim.2018.02.008

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In Leary M. R., Hoyle R. H. (eds.), Handbook of individual differences in social behavior. New York, NY: Guilford Press, 369–381.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass, 14*(11), e12633. https://doi.org/10.1111/phc3.12633

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology: Interdisciplinary and Applied, 55*(1), 33–38, https://doi.org/10.1080/00223980.1963.9916596

Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, R., … , & Zito, T. (2017). Sustainable computational science: the ReScience initiative. *PeerJ Computer Science*, 3, e142. https://doi.org/10.7717/peerj-cs.142

Sandve, G. K., Nekrutenko, A., Taylor. J., & Hovig, E. (2013) Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Scheel, A. M. (2019, March 12). Positive result rates in psychology: Registered reports compared to the conventional literature. ZPID (Leibniz Institute for Psychology Information). https://doi.org/10.23668/psycharchives.2390

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325–1346. http://dx.doi.org/10.1037/bul0000169

Steinle, F. (2016). *Stability and replication of experimental results*. In Atmanspacher H., Maasen S. (eds.), *Reproducibility: Principles, problems, practices, and prospects*, Wiley: Hoboken, NJ, 39–64.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association, 54*(285), 30–34.

Sterling T. D., Rosenbaum W. L., Weinkam J. J. (1995), Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*, 108–112. https://doi.org/10.1080/00031305.1995.10476125

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, 115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Szucs, D., Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, 1(1), Art. 3. http://doi.org/10.1525/collabra.13

Zeigler, B. P., Muzy, A., & Kofman, E. (2019). *Theory of modeling and simulation: Discrete event and iterative system computational foundations* (3rd ed.). London: Academic Press.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726