Paper submitted to special issue "Curiosity, Imagination and Surprise" of Research in the History of Economic Thought and Methodology, January 2018

Re-Thinking Reproducibility as a Criterion for Research Quality

Sabina Leonelli
Exeter Centre for the Study of the Life Sciences & Department of Sociology, Philosophy and Anthropology
University of Exeter
s.leonelli@exeter.ac.uk

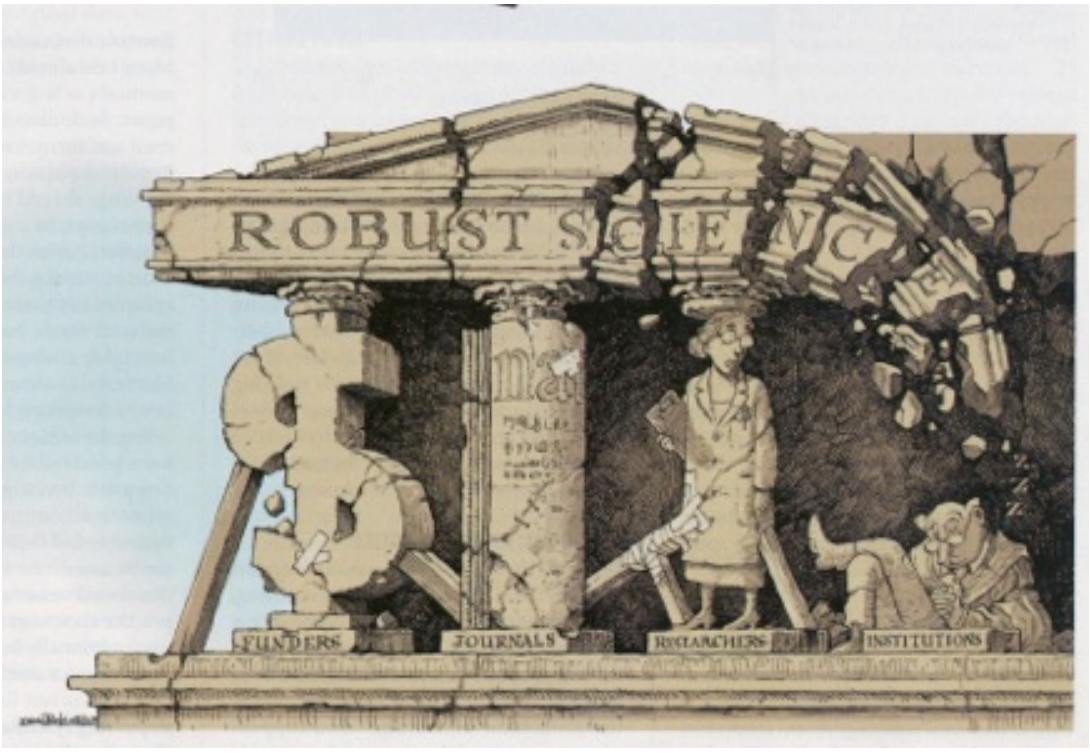
Abstract: A heated debate surrounds the significance of reproducibility as an indicator for research quality and reliability, with many commentators linking a "crisis of reproducibility" to the rise of fraudulent, careless and unreliable practices of knowledge production. Through the analysis of discourse and practices across research fields, I point out that reproducibility is not only interpreted in different ways, but also serves a variety of epistemic functions depending on the research at hand. Given such variation, I argue that the uncritical pursuit of reproducibility as an overarching epistemic value is misleading and potentially damaging to scientific advancement. Requirements for reproducibility, however they are interpreted, are one of many available means to secure reliable research outcomes. Furthermore, there are cases where the focus on enhancing reproducibility turns out not to foster high-quality research. Scientific communities and Open Science advocates should learn from inferential reasoning from irreproducible data, and promote incentives for all researchers to explicitly and publicly discuss (1) their methodological commitments, (2) the ways in which they learn from mistakes and problems in everyday practice, and (3) the strategies they use to choose which research component of any project needs to be preserved in the long term, and how.

Keywords: reproducibility; research methods; data practices; research methods; pluralism.

Introduction: The Reproducibility Crisis

The reproducibility of research results is often flagged as a priority in scientific research, a fundamental form of validation for data and a major motivation for open data policies (Royal Society 2012, Science International 2015). The epistemic significance attributed to reproducibility has recently become poignant as a result of the failure of studies aiming to reproduce the results of well-established, published research in the fields of molecular biology (Ioannidis et al 2012), experimental psychology (Open Science Collaboration 2015; Baker 2015) and clinical medicine (Prinz et al 2011; Begley and Ellis 2012). The findings of these "reproducibility tests" have been published and commented upon in prominent international journals within science and beyond (e.g. as the main feature in a 2013 issue of *The Economist),* causing widespread worry and uproar. Failure to reproduce results was taken to indicate, at best, problems in the design and methods used by the researchers in question, and at worst, a fundamental lack of credibility for the knowledge thereby obtained – a position reminiscent of Karl Popper's pronouncement that "non-reproducible single occurrences are of no significance to science" (1959, 64). In 2015, *Nature* published a full special issue on the "challenges of irreproducible research" (http://www.nature.com/news/reproducibility-1.17552), which included a cartoon showing the ruins of the temple of "robust science" [figure 1].

Figure 1. From Nature, September 15, 2015. Reproduced by courtesy of Nature Publishing Group.



Many commentators pointed to the lack of reproducibility as a signal that, due to increasing administrative, social and financial pressures, scientists are becoming sloppier and eager to publish as soon as feasible, without taking adequate time and effort to verify the reliability of their results. This encourages the generation of low-quality research outcomes, which should not be relied upon as sources of knowledge. This so-called "crisis of reproducibility" has also been linked to: ineffectual practices of quality control and self-correction within journals, resulting in the failure to pick up serious errors and unreliable results at the refereeing stage (Allen et al, 2016); a general increase in the complexity and scale of experiments and related statistical analyses of results, with increasing specialization and division of labor resulting in lack of clarity around who is actually responsible for quality checks (Leonelli 2017a); widespread cognitive bias among researchers, who end up reinforcing each other's preexisting beliefs and setting up research plans and validation procedures accordingly (Bem, Utts and Johnson 2011); questionable uses of statistical techniques to smoothen bias and exclude uncomfortable results (e.g. the practice of p-hacking and selective reporting; Simmons, Nelson and Simonsohn 2011, Fanelli 2012); and lack of transparency and effective critical debate around research methods and data, which makes it impossible for researchers outside any given project to scrutinize its results (Science International 2015). Among the potential consequences of the crisis, commentators list an increasing mistrust in the credibility of scientific findings by policy-makers and citizens, the squandering of private and public investments in the search for new technologies and medical treatments, and permanent damage to the integrity and ethos of the scientific enterprise (e.g. KNAW 2018). It is hardly possible to imagine higher stakes than these for the world of science. Its future existence and social role seems to hinge on the ability of researchers and scientific institutions to respond to the crisis, thus averting a complete loss of trust in scientific expertise by civil society.

I find many of the critical reflections on the challenges presented by the current scale, organization and methods of much experimental research to be timely and compelling, and I agree that traditional models of scientific publishing and public scrutiny are no longer fit for purpose within the increasingly globalized, costly, technology-dependent and multi-disciplinary landscape of contemporary research (Leonelli 2016, 2017a/b). However, in this paper I take issue with the widespread reference to reproducibility as an overarching epistemic value for science and a good proxy measure for the quality and reliability of research results. Reproducibility comes in a variety of forms geared to different methods and goals in science. While most commentators focus on the use of reproducibility as the best strategy to achieve inter-subjectively reliable outcomes, I shall argue that (1) such convergence of results can be obtained in the absence of reproduction, and (2) a different, yet crucial function of reproducibility consists in helping researchers to identify relevant variants in

the first place.

My analysis is organized as follows. Taking inspiration from some of the scientific and philosophical discussions on reproducibility, I discuss how the meaning of this term can shift dramatically depending on what it is applied to, and how. I then review how such variability plays out in practice by identifying six broad types of research associated with different interpretations of reproducibility. I consider how this variation is linked to the degree of control that researchers are able and willing to exercise on their materials and on the environment in which their investigations take place, as well as to the extent to which they rely on statistical methods in assessing the validity of the evidence being produced. In studies where control over environmental variants is only partially achieved, for instance, reproduction resulting in different outcomes is perceived as highly valuable, since it can signal hitherto unknown sources of variation or define the scope of the hypothesis being tested. By contrast, in studies that are carried out in highly idiosyncratic environmental conditions and/or on perishable and rare samples which do not lend themselves to statistical analysis, it is the very uniqueness and irreproducibility of research conditions that makes the resulting data valuable as sources of evidence. In such cases, a focus on enhancing reproducibility turns out not to be the best way to foster high-quality, robust research outcomes. Rather, it is the well-informed analysis of how reliable and meaningful data are obtained through irreproducible research practices that increases the sophistication of research methods and of the ways in which they are documented and disseminated. This often includes an emphasis on the triangulation of results obtained through a variety of methods and under diverse circumstances – thereby taking convergence among findings coming from different sources as signaling the robustness and reliability of a particular inference.[1]

I conclude that the uncritical pursuit of a narrow interpretation of reproducibility as an overarching epistemic value for research, endowed with the power to demarcate science from non-science, is misleading and potentially damaging to scientific advancement. Reproducibility is not only interpreted in many differently ways, but also serves a variety of epistemic functions depending on the field and type of research at hand. Requirements for reproducibility, however they are interpreted, are thus only one of the available means to secure reliable research outcomes. In the final section of the paper, I highlight the implications of this argument for scientific communities and Open Science advocates.

[1] While maintaining a focus on repetition as a conclusive test, Allan Franklin (1986) discusses other several strategies for enhancing trust in experimental results. Chapman and Wylie (2016) provide an excellent discussion of triangulation as a key strategy for testing the robustness of inferential reasoning in the historical sciences.

What Is Reproducibility?

Within both the natural sciences, the philosophy of science and science studies there is a wide variety of interpretations of the meaning and practical implications of the term 'reproducibility' - and of associated terms such as replicability and repeatability. This should not be surprising, given the vast array of goals and concerns related to reproducibility that exist across different fields and approaches to research. Within applied domains such as pharmacology and clinical medicine, reproducibility is widely regarded as a way to assess the safety of very specific knowledge claims and related products (such as drugs), and the circumstances in which such safety can be guaranteed. If an experiment fails to yield the same results when tried on different populations, for instance, reproducibility can be used as a tool to investigate the range of validity of the inferences being tested. Reproducibility can also be invoked for other goals, however, such as: the debugging of software from errors in computer science and informatics; the exploration of how and why methods tried on different materials, or on the same materials at different times or in different places, may yield different outcomes – as is often the case in the life sciences; or the investigation of the effect of researchers' own biases, assumptions and interests on the outcomes of research – a crucial issue particularly in experimental psychology and behavioural economics.[2] As noted by Hans Radder:

> "many philosophers assume, implicitly or explicitly, that successful experiments are or should be reproducible. However, since "experiment" is a general term for what in fact is a rather complex process, the precise meaning of this assumption is not clear. To clarify the notion of reproducibility we need to address the following question: reproducibility *of what* and *by whom*?" (Radder 1996, 16).

This question becomes ever more vexing when considering research that does not rely on experimental techniques but rather on observational data, including both quantitative approaches such as surveys and qualitative methods such as ethnography. This landscape can be very confusing to analysts who are trying to relate field-specific discussions of reproducibility to each other, particularly since commentators tend to use the same terms in different ways – and often to discuss different problems.

Rather than attempting to review in detail this myriad of terminological differences, I focus on two central concerns that all commentators seem to agree

[2] Cases of reproducibility failure in economics are less widely discussed, but are no less interesting. Julie Nelson re-analysed existing studies suggesting strong evidence for women being more risk-averse than men, and showed how they cherry-picked findings, used bogus statistical methods and visualized data in ways that suggested indefensible claims (for instance, she showed that as soon as you visualize the data showing wide variation among responses in both men and women, the effect diminishes and it becomes clear that risk-aversion is not easily correlated to gender differences between individuals; Nelson 2016).

on.[3] One is the distinction between the ability to obtain the same results through the same methods, and the ability to obtain the same results through an array of different methods. Radder refers to the former as *reproducibility* and to the latter as *replicability*, thus associating the term reproducibility with the requirement to successfully repeat the whole of a research procedure in order to ascertain the reliability of its outcomes (Radder 1996, 2012). This in turn requires a very high degree of control over sources of variability that may affect not only the outcomes, but also the very performance of an investigation (such as, for instance, whether researchers can reliably expect a certain set of reactions from a given organism or material under investigation to a given environmental stimulus).

The second key concern is what is meant by 'the same outcomes'. What commentators and researchers view as outcomes can encompass objects and processes at different levels of abstraction and produced at different stages of research. These range from direct *measurements* generated by experimental apparatus (what Ian Hacking calls "marks" and many others refer to as "raw data", to signal the relative independence between the measurements and subsequent interpretations; Hacking 1995, Leonelli 2016) to *patterns* extracted from such measurements in the form of data models, *causal generalisations* derived from data analysis in the form of knowledge statements, or *forms of intervention* such as those used to create a new chemical substance or modify the structure of a cell. Whether outcomes are considered to be data, models, interventions or knowledge statements matters enormously to the interpretation of claims around reproducibility, since the reproduction of the same outcomes can be interpreted as the requirement to obtain precisely the same dataset multiple times or, more liberally, as being able to derive the same patterns and/or generalisations from that dataset multiple times and across different circumstances (even if the data themselves are not precisely the same).

Remarkably, despite the variable interpretations of the term that are immediately visible through such basic considerations, the framing of reproducibility used in top science journals today remains narrowly linked to a particular understanding of what good research should look like. Most typically, reproducibility is associated with experimental research methods that yield numerical outcomes. Existing scholarship in the philosophy of measurement and experimentation points to at least three reasons for this: (1) experimentation aims to achieve a measure of control over sources of variability that may affect the outcomes of research, and the higher the degree of control that researchers have over materials and set-up, the higher the chance that repeating experiments will yield the same outcomes (Radder 2012); (2) quantitative outcomes are easier to compare with each other and aggregate into patterns than other types of data, thus enabling what looks like a simple and immediate comparison between research outcomes (Porter 1995); (3) numbers lend themselves to computationally and statistically driven forms of analysis and validation, which are often assumed to lend a measure of mechanical objectivity

[3] An excellent review of the various ways in which the term replicability is used in the life sciences, with explicit reference to the distinctive phases of research projects, is provided by Shavit and Ellison (2017, 9-14).

to the evaluation of research outcomes (Daston and Galison 1992, Boumans 2015). As these authors have recognised and I discuss in the next section, however, this approach to scientific research is not the only one yielding useful and reliable results, nor should it be assumed to be. The interpretation of reproducibility associated with this mode of research works better for certain methods, stages and goals of inquiry than for others, thus proving to be inadequate as an overarching criterion for what reliable, high-quality research needs to look like.

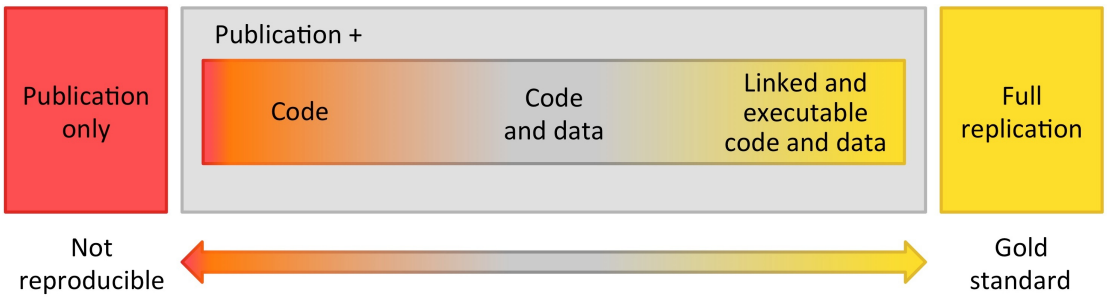
Reproducibility in Scientific Practice

I here reflect on how reproducibility is interpreted in relation to methodologies employed in research, the relation between these interpretations and the degree of control that researchers have on their environment, and the extent to which they rely on statistical inference. This is not meant to be an exhaustive list, nor is it meant to apply accurately to all existing instances of the examples being mentioned. Rather, it is meant to convey a sense of the patterns and methodological standards characterizing different parts of the scientific world, and the diversity of assumptions and practices that can be associated with the ideal of reproducibility in research.

1 Computational Reproducibility

Radder's definition of reproducibility is perhaps best matched by the practices of software development and validation used in computer science and informatics, where researchers focus on finding and resolving mistakes and bugs in data analysis by running the same data through a given set of algorithms over and over again. Here it is not only possible, but actually necessary to reduce any source of variation other than in the software or statistical tools being used to a minimum, so as to spot whatever error may have infiltrated the system. As discussed in a recent volume aimed to provide an overview of reproducibility techniques, a well-established way to capture this is the idea of "computational reproducibility": "A research project is *computationally reproducible* if a second investigator [..] can recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions." (Kitzes 2016, 12). A paper published in *Science* in 2011 usefully links this interpretation of reproducibility with the use of minimal standards for publishing meta-data within computational science. As shown in their graphical illustration of what they call "reproducibility spectrum", captured in figure 2 below, this view of reproducibility does not involve replicating the circumstances of data production, but rather being able to obtain the same outcomes when running a given dataset through the same algorithms.

Figure 2. Reproducibility spectrum according to Peng (2011), redrawn by Michel Durinx.



2 Direct Experimental Reproducibility: Standardised Experiments

The circumstances of data production are, by contrast, a primary concern for experimentalists, and indeed computational reproducibility is hardly applicable to experimental research, where one is unavoidably confronted with variation. At the same time, there is significant disagreement among experimental fields concerning what is seen as a desirable and realistic degree of standardisation and control over environmental variables. In clinical trials aimed to test the safety and efficacy of given drugs vis-à-vis human populations, the degree of controls and standardization being implemented is among the most impressive and tightly scrutinized within the biomedical realm. Experiments conducted in particle accelerators in physics are similarly controlled, the focus on one centralized experimental apparatus being particularly helpful in establishing a fixed framework within which experiments can be successfully repeated. The concept of reproducibility pursued in such settings is what I shall hereafter refer to as 'direct reproducibility': the ability to obtain the same results through the repeated application of the same research methods. This is as close to computational reproducibility as it can get within the experimental sciences, though nobody – particularly in clinical science - would expect a complete and exact match between the results of different runs of the same experiment. Rather, what is expected is a strong similarity in the datasets and a match in the patterns inferred from the data (KNAW 2018). This is typically evaluated through recourse to statistical inference, thus privileging statistics as a key validating tool for reasoning from evidence.[4]

3 Scoping, Indirect and Hypothetical Reproducibility: Semi-Standardised Experiments

Most experiments are carried out under conditions that are less tightly controlled than clinical research. Such experiments are best described as semi-

[4] This is what Julian Reiss called "experimental paradigm" (Reiss 2015).

standardised: that is, where methods, set-up and materials used have been construed with ingenuity in order to yield very specific outcomes, and yet some significant parts of the set-up necessarily elude the controls set up by experimenters. Typically, it is those 'non-standardised' components that yield the most valuable epistemic insights for researchers, by generating confounding results and enabling the exploration of new phenomena.[5] One example is research on model organisms, where researchers have spent decades engineering the organisms to conform to specific morphological, developmental and behavioural standards, and yet the behaviour and structure of the organisms remains in part unpredictable (a veritable 'sample of nature'; Leonelli 2013) and highly susceptible to subtle shifts in environmental circumstances, ranging from nutrition to lighting and temperature (Ankeny and Leonelli 2011, Reardon 2016).[6] Another example is psychological experiments on social groups selected because conforming to given physical, social and behavioural criteria, and yet presenting unforeseen sources of variability of potential relevance to the outcomes being generated. A third notable case is that of brain scans and other types of brain imaging in neuroscience, which are often affected by imperceptible shifts in the subject's mood and metabolism despite the tight control exercised by researchers on external stimuli (Turner and De Haan 2017).

Within this type of research, it is common to find complaints about how contextual differences between laboratory settings, research objects and other environmental circumstances compromise the extent to which researchers can aim for reproducibility. And yet, many researchers working under these conditions do not aim for direct reproducibility. Some run experimental reproductions to spot sources of variation that may prove significant when interpreting the data at hand – for instance, in the case of results obtained on model organisms, when establishing the extent to which a given outcome can be reliably imputed to organisms beyond those originally used in the study. This is an interpretation that I will call *scoping reproducibility*.[7]

Others prefer other yet interpretations. One is *indirect reproducibility* that focuses on obtaining similar results from the performance of different experiment (what Radder called replicability), and constitutes a useful validation tool to see whether results produced under variable circumstances converge or not. Another is what Felipe Romero calls *hypothetical reproducibility* (or

[5] Mary Morgan (2012a, 296) introduced the idea of confoundment as resulting from an experimental outcome challenging researchers' existing knowledge of the world, and thus leading to the discovery of new phenomena. This is contrasted to the surprise triggered by model experiments, where discovery concerns the world of the model, rather than the world itself.

[6] A recent Nature article complained of the difficulties in controlling for environmental variability within studies on mice: "Mice are sensitive to minor changes in food, bedding and light exposure. It's no secret that therapies that look promising in mice rarely work in people. But too often, experimental treatments that succeed in one mouse population do not even work in other mice, suggesting that many rodent studies may be flawed from the start. Researchers rarely report on subtle environmental factors such as their mice's food, bedding or exposure to light; as a result, conditions vary widely across labs despite an enormous body of research showing that these factors can significantly affect the animals' biology. "It's sort of surprising how many people are surprised by the extent of the variation" between mice that receive different care, says Cory Brayton, a pathologist at Johns Hopkins University in Baltimore, Maryland." (Reardon 2016: 264)

[7] I thank Mary Morgan, Hans Radder and Stephan Güttinger for discussions on this point.

'conceptual replication' in psychologists' jargon; Romero 2017): the attempt to obtain outcomes that match those predicted as implications of previous findings, thereby confirming the reliability of the previous findings. Both these interpretations of reproducibility lean on the idea that convergence across multiple lines of evidence, even when they are produced in different ways, is a mark of reliable research (a 'tangle of support' in Chapman and Wylie's terms, 2016).

4 Reproducible Expertise: Non-Standard Experiments and Research on Rare Materials

There are also experiments where control over environmental variability is extremely limited, and standardisation very low. These are cases where experimenters are studying new objects or phenomena (new organisms for instance) and/or employing newly devised, unique instruments that are precisely tailored to the inquiry at hand.[8] Researchers then focus less on controls and more on developing robust ways of evaluating the effects of their interventions and the relation between those effects and the experimental circumstances at the time in which data were collected. Direct or indirect reproducibility are not helpful concepts within this type of research. Notably, however, the idea of reproducibility does not completely disappear, with researchers emphasising the reproducibility of the expertise – the specific skills and interpretive abilities – underpinning the conduct of research over the reproducibility of the outcomes. I will refer to this interpretation of reproducibility as *reproducible expertise*, and define it as the expectation that any skilled experimenter working with the same methods and the same type of materials at that particular time and place would produce similar results.

Appeals to reproducible expertise are also characteristic of research on materials that are rare, unique, perishable and/or inaccessible, such as depletable samples stored in biobanks; unique specimens, such as specific botanical finds or archaeological remains; or materials that are hard or expensive to access, such as very costly strains of transgenic mice). These materials are not amenable to repeated investigation as required by the direct and indirect forms of reproducibility. This does not constitute an obstacle to using such materials for research, since the uniqueness and irreproducibility of the materials is arguably what makes the resulting data particularly useful as evidence. The onus of reproducibility shifts instead to the credibility and skills of the investigators entrusted with handling these materials. Apposite methodologies have been developed to cope with the impossibility to directly replicate the findings, including vetted access, cross-samples research and the centralisation of research in locations where several researchers can work together and check each other's work and ensure its reliability for those with no access to the same material sources.

[8] Ed Ramsden, Rachel Ankeny, Nicole Nelson and I discussed models of organisms that include environmental features and are uniquely tailored to the specific goals of a given inquiry as 'situated models' (Ankeny et al 2014).

5 Reproducible Observation: Non-experimental case description

A tremendous amount of research in the medical, historical and social sciences does not rest on experimentation, but rather on observational techniques such as surveys, descriptions and case reports documenting unique circumstances (Morgan 2012b). Such research is, again, not replicable in the sense of direct, indirect or even hypothetical reproducibility (Boumans 2015) and yet, as in the case of non-standard experiments, there is an emphasis on reproducibility of observation - the expectation being that any skilled researcher placed in the same time and place would pick out, if not the same data, at least similar patterns. In other words, one can learn to observe in very specific, reproducible ways. Examples are the practices of comparative multi-sited ethnography, where researchers are trained to observe similar phenomena across very different circumstances; structured interviewing, where researchers devise a relatively rigid framing for their interactions with informants; and diagnosis based on radiographies, resonance scans and other medical imaging techniques, where skilled observation by expert physicians is crucial to extracting meaningful and reliable information.

6 Irreproducible Research: Participant Observation

At the other end of the spectrum from computational reproducibility there is research where the idea of reproducibility has been rejected in favor of an embrace of the subjectivity and unavoidable context-dependence of research outcomes. Researchers working with highly idiosyncratic, situated findings are well-aware that they cannot rely on reproducibility as an epistemic criterion for data quality and validity. They therefore devote considerable care to documenting data production processes and strategizing about data preservation and dissemination. In other words, they prioritize sophisticated strategies for enhancing the accountability of their methods and data management strategies, as well as the long-term preservation of the instruments, techniques and materials through which results were generated. The very fact that different observers have different viewpoints and produce different data and interpretations is here used as a starting point for assessing and validating research results. Ethnographic work in anthropology, for instance, has developed methods to account for the fact that data are likely to change depending on time, place, subjects as well as researchers' moods, experiences and interests. Key among such methods is the principle of reflexivity, which requires researchers to give as comprehensive a view of their personal circumstances during research as possible, so as to enable readers to evaluate how the focus of attention, emotional state and existing commitments of the researchers at the time of the investigation may have affected their results. Much can be learnt within the quantitative sciences from this approach, which explicitly and publicly discusses the methodological commitments and processual nature of the research (including the ways in which investigators

change their work in response to shifting circumstances, problems and mistakes in everyday practice), and makes a virtue out of the unavoidable variation among studies carried out at different times, by different groups and in different places.

Table 1. Synoptic view of types of research design/methods and related understanding of reproducibility discussed in section 3.

Type of research	**Example**	**Degree of control on environment**	**Reliance on statistics as inferential tool**	**Reproducible in which sense?**
Software development	Computer engineering, informatics	*Total*	High	*Computational R*: Obtain same results from the same data.
Standardised experiments	Clinical trials, environmental safety controls	Very high	High	*Direct R:* Obtain same results from different runs of the same experiment.
Semi-standardised experiments	Behavioural economics, experimental psychology, research on model organisms	Limited	Variable	*Scoping R:* Use differences in results to identify relevant variation. *Indirect R:* Obtain same results from different experiments. *Hypothetical R:* corroborate results implied by previous findings.
Non-standard experiments & research based on rare, unique, perishable, inaccessible materials	Research on experimental organisms, archeology, paleontology, history	Low	Low	*Reproducible Expertise:* Any skilled experimenter working with same methods and materials would produce similar results
Non-experimental case description	Case reports in medicine, (types of) multi-sited ethnography	None	Low	*Reproducible Observation:* Any skilled observer would pick out similar patterns
Participant observation	Ethology, participant observation in anthropology	None	None	*Irreproducible Observation:* different observers are assumed to have different viewpoints and produce

				different data and interpretations

Beyond the Ideal of Direct Reproducibility

The idea that reproducibility, interpreted univocally as the reproduction of the same outcomes by the same methods, cannot work as an overarching criterion for evaluating the quality and reliability of research is by no means new. A well-known critic of this view is Harry Collins, whose seminal book *Changing Order* argued that the vast majority of experiments are never replicated, and even in the few cases where replication is attempted, it is not always successful (Collins 1985). It is nevertheless interesting to note that while critical of the use of reproducibility as a descriptive and evaluative tool, Collins is sympathetic to the use of reproducibility as a regulatory ideal for science: "even though replication is beset by problems, one must stick to it as the fundamental way of showing that scientific results are secure. Science is a matter of having the right aspirations even if they cannot be fulfilled" (Collins 2017).

I disagree with this view. Seeing reproducibility as a key aspiration for science brings us back to the Popperian view that reproducibility, at least heuristically, demarcates good from bad science. There are good reasons to resist the temptation to impute such decisive epistemic power to reproducibility. As I discussed, this ideal is applied very differently depending on research fields, materials and goals. Nevertheless, reproducibility as currently discussed in science and science policy is often linked to the expectation that researchers can and should be able to exercise a high level of control over the circumstances, environment and materials employed in a study. Insisting on this interpretation of reproducibility, particularly within funding and assessment structures, can push researchers to place less emphasis on carefully reporting the more idiosyncratic aspects of their research, and instead focus on producing general protocols that do not linger on the specific characteristics of their local situation. It can also be interpreted as incentivising researchers to focus less attention on the variation characterizing their results, and the extent to which such variation can affect the reliability and scope of their conclusions. This situation carries significant epistemic risks. As reported by *Nature*, "researchers rarely report on subtle environmental factors such as their mice's food, bedding or exposure to light; as a result, conditions vary widely across labs despite an enormous body of research showing that these factors can significantly affect the animals' biology" (Reardon 2016). A contrasting and more productive attitude is to repeat experiments not as a way to reproduce results, but rather as a way to uncover sources of variation and explore their significance (what I already discussed under the label of 'scoping reproducibility').

Generally, the emphasis on a narrow interpretation of reproducibility is linked with a devaluing of the role of expertise and embodied knowledge in data production, processing and assessment. These are instead highly valued and well-accounted for in qualitative research traditions that focus on the critical evaluation of data in light of the situatedness of human perspective and the local

nature of research methods and materials. This results in much useful work on the meta-data that should accompany and guide the analysis of any given dataset (e.g. Zahle 2018), and the insights that can be gained when obtaining diverse outcomes from the same research process. It also fosters research on triangulation techniques aiming to compare and integrate results coming from different traditions, locations, sources and methodologies, which in turn facilitates testing whether any given inference is robust in the face of different lines of evidence.

Given these issues, it is important to ask why direct reproducibility proves so attractive as an ideal to which research should aspire. An insightful explanation for the epistemic power attributed to this approach is provided by another critic of this view, John Norton. His skepticism is grounded partly in the observation that "a failure of replication may not impugn a credible experimental result; and a successful replication can fail to vindicate an incredible experimental result"; and partly in the view that whether or not an experiment may be viewed as successfully replicated is determined by whichever background facts researchers appeal to when evaluating the experiment and its results. As Norton notes, "commonly, these background facts do support successful replication as a good evidential guide and this has fostered the illusion of a deeper, exceptionless principle" (Norton 2015). In other words, researchers that pursue direct reproducibility have been so successful at engineering cohesive experimental systems that highly controlled experiments have come to exemplify the very best of research practices, in ways that do no justice to other research methods, and particularly to qualitative traditions.

Conclusion: The Epistemic Value of Irreproducible Research

We have seen how direct reproducibility works best in research environments characterized by a high standardization of methods and materials, a high degree of control over environmental variability and reliable and relevant methods of statistical inference. It should not be surprising that research that strays from these conditions – such as exploratory, non-standard research carried out on unique samples and under highly variable environmental conditions - has trouble conforming to this interpretation of reproducibility, and I have suggested here that such conformity is neither fruitful nor desirable. In non-standard types of inquiry, researchers typically recognize that direct reproducibility cannot function as an epistemic criterion for research quality, and instead devote care and critical thinking on documenting data production processes, examining the variation among their materials and environmental conditions, and strategizing about data preservation and dissemination. Within qualitative research traditions, explicitly side-stepping the ideal of (direct) reproducibility has helped researchers to improve the reliability and accountability of their research practices and data.

This is rarely acknowledged in scientific debates, but there are some interesting

exceptions, particularly as the reproducibility debate is reaching maturity and the initial polemical tones are being replaced by more sophisticated reasoning. A recent example is a piece written by the editor of *Science* journals in January 2018, where he defends the steps taken within the journals to improve reproducibility, but also notes that

> "Another approach to assess reproducibility involved an experimental program that attempted to replicate selected findings in cancer biology by groups not involved with the original studies (see https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology). Although some findings were largely reproduced, in at least one case (which was published in *Science*), the key finding was not. Yet, the initial results have been utilized and extended in published studies from several other laboratories. This case reinforces the notion that reproducibility, certainly in cancer biology, is quite nuanced, and considerable care must be taken in evaluating both initial reports and reported attempts at extension and replication. Clear description of experimental details is essential to facilitate these efforts. The increased use of preprint servers such as bioRxiv by the biological and biomedical communities may play a role in facilitating communication of successful and unsuccessful replication results." (Berg 2018)

This is an important step towards acknowledging that the circumstances of reproducibility differs across research fields. It falls short of an acknowledgement that reproducibility itself can come in different guises, and sometimes be instrumental to a different epistemic purpose or altogether irrelevant. Here again, the editorial equates science with experimentation, thereby excluding non-experimental sciences from consideration.

What is a credible alternative to the current tendency to rely on narrow forms of reproducibility as a fail-proof criterion for what should be trusted as good science? One option is to recognise the importance of reproducibility as a strategy to interrogate variation across results, rather than to validate the outcomes of research. Another option is to focus on the idea of convergence or triangulation of results, as discussed above. Perhaps the strongest advocate of that view in philosophy has been Hacking, when arguing that scientific results can be trusted when the elements of laboratory science are brought into mutual consistency and support, match each other and become "mutually self-vindicating" (Hacking 1992, 56).[9] A third alternative is to tailor reproducibility requirements to the circumstances and goals of any specific project or area of research, while at the same time learning from situations where insisting on reproducible results is neither realistic nor significant to assessing the quality of a given study. Researchers who cannot rely on reproducibility as an epistemic

[9] The elements that Hacking singles out include "(1) ideas: questions, background knowledge, systematic theory, topical hypotheses, and modeling of the apparatus; (2) things: target, source of modification, detectors, tools, and data generators; and (3) marks and the manipulation of marks: data, data assessment, data reduction, data analysis, and interpretation" (Hacking 1992, 56) – thus potentially embracing research beyond the experimental sciences.

criterion for quality and validity tend to devote more care and critical thinking to accounting for research methods and circumstances – for instance by documenting data production processes (including the mood and circumstances of individual researchers) and strategizing about the preservation and dissemination of related instruments, protocols and materials.

Indeed, many scientific communities and Open Science advocates focus their efforts on fostering public discussion of researchers' methodological commitments and everyday practices. In consultation with learned societies and experts in qualitative research, funding bodies and research organisations should mirror these efforts by providing researchers with incentives and guidelines to explicitly discuss not only their methods, but also the ways in which they learn from unexpected and incongruent findings. Furthermore, overt discussions should include strategies to preserve research components and materials in the longer term, particularly in situations where hard choices need to be made about which instrument, software, material and meta-data it is actually possible (and realistic) to store, and how. A frank discussion of the extent to which everyday research practice deviates from idealised reproducibility standards is not only possible, but crucial to the exercise of critical scrutiny and constant questioning of established knowledge that Popper himself viewed as markers of good science. The goal of such open exchange should not be to punish research approaches where computational or direct reproducibility is not possible or relevant, but rather to acknowledge the strengths and weaknesses of different ways of validating results, and learn as much as possible from the methodological precepts that guide different parts of science.

Acknowledgments

This paper – like much of my work over the last ten years – owes much to Mary Morgan's generous advice, and her ability to identify and probe the core and significance of arguments and questions. I also gratefully acknowledge insightful comments by Hans Radder, Stephan Güttinger, Niccolo Tempini, two anonymous referees, and participants to the conference "Curiosity, Imagination and Surprise" (Utrecht, September 2017), particularly the host Marcel Boumans. This research was funded by the European Research Council grant award 335925 ("The Epistemology of Data-Intensive Science"), the Australian Research Council Discovery Project "Organisms and Us" and the UK Economic and Social Research Council award ES/P011489/1.

References

Allison, D. B. et al (2016). A tragedy of errors. *Nature*, *530*, 27–30.

Ankeny, R.A., Leonelli, S., Nelson, N. and Ramsden, E. (2014) Making Organisms Model Humans: Situated Models in Alcohol Research. *Science in Context* 27(3): 485-509.

Baker, L (2015) Over Half of Psychology Studies Fail Reproducibility Test. *Nature* doi:10.1038/nature.2015.18248

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. http://doi.org/10.1038/483531a

Bem, D. J., Utts, J, Johnson W. O. (2011) Must Psychologists Change the Way They Analyze Their Data? *Journal of Personality and Social Psychology* 101 (4): 716–719.

Berg, J (2018) Progress on reproducibility. *Science* 359(6371): 9 DOI: 10.1126/science.aar8654

Boumans, M (2015). *Science outside the Lab*. Oxford University Press.

Collins, H. (2017) in Justin Kitzes, Daniel Turek, Fatma Deniz (Eds.) The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. University of California Press.

Collins, H. (1985). *Changing Order: Replication and Induction in Scientific Practice*. London: Sage Publications.

Daston, L and Galison, P (1992) "The Image of Objectivity." *Representations* 40: 81-128.

Earp, B.D. and Trafimow, D. (2015) Replication, falsification, and the crisis of confidence in social psychology, *Frontiers in Psychology*, 6(621), 1-11 (doi: 10.3389/fpsyg.2015.00621).

Fanelli, D. (2012) "Negative Results are Disappearing from Most Disciplines and Countries." *Scientometrics* 90 (3): 891–904.

Franklin, A. (1986) *The Neglect of Experiment*. Cambridge, UK: Cambridge University Press.

Hacking, I. (1983). *Representing and Intervening.* Cambridge, UK: Cambridge University Press.

Hacking, I. (1995). Introduction, in: J.Z. Buchwald (ed.) *Scientific Practice: Theories and Stories of Doing Physics.* Chicago: University of Chicago Press, 1-9.

Ioannidis et al (2010) Nature Genetics 41 on repeatability of published microarray gene expression analyses – did not detail software used

Kitzes, J (2017) in Justin Kitzes, Daniel Turek, Fatma Deniz (Eds.) The Practice of *Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. University of California Press.

KNAW (2018). *Replication studies – Improving reproducibility in the empirical sciences*, Amsterdam, KNAW.

Leonelli, S. (2016) *Data-Centric Biology: A Philosophical Study*. Chicago, Il: Chicago University Press.

Leonelli, S. (2017a) Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal* 16(32): 1-11. DOI: https://doi.org/10.5334/dsj-2017-030

Leonelli, S. (2017b) *Incentives and Rewards to Engage in Open Science Activities.* Report for the Mutual Learning Exercise of the European Commission: Open Science – Altmetrics and Rewards. https://rio.jrc.ec.europa.eu/en/library/mutual-learning-exercise-open-science-%E2%80%93-altmetrics-and-rewards-incentives-and-rewards-engage

Leonelli, S. (2013) 'Model Organism'. In: Dubitzky, W., Wolkenhauer, O., Cho K-H., Yokota, H. (Eds.) *Encyclopaedia of Systems Biology*. Springer.

Morgan, M. S. (2012a) *The World in the Model: How Economists Work and Think.* Cambridge, UK: Cambridge University Press.

Morgan, M. S. (2012b) Case Studies: One Observation or Many? Justification or Discovery? *Philosophy of Science* 79:5, 667-77.

Nelson, J. (2016) Not-So-Strong Evidence for Gender Differences in Risk Taking, *Feminist Economics* 22(2), 2016, pp. 114-142.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, 349(6251), 1-8.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science* 334, 1226–1227.

Popper, K. (1959 [2002]) *The Logic of Scientific Discovery*. London: Routledge.

Porter, T. M. (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery*, *10*(9), 712.

Pulverer, B (2015) Reproducibility Blues. EMBO. http://emboj.embopress.org/content/34/22/2721.full

Radder, H. (2012[1984/1988]). *The Material Realization of Science. From Habermas to Experimentation and Referential Realism*, revised edition, with a new postscript. Dordrecht: Springer.

Radder, H. (1996). *In and About the World: Philosophical Studies of Science and Technology*. State University of New York Press.

Reardon, S (2016) A mouse's house may ruin experiments: Environmental factors lie behind many irreproducible rodent experiments. *Nature* 530, 264.

Reiss, J. (2016) *Causation, Evidence and Inference*. London: Routledge.

Romero, F. (2017) Novelty vs Replicability: Virtues and Vices in the Reward System of Science. *Philosophy of Science*.

Royal Society (2012) *Science as an Open Enterprise*. Accessed January 14 2018. http://royalsociety.org/policy/projects/science-public-enterprise/report/

Science International (2015) *Open data in a big data world*. Last Accessed 19 January 2018, URL: https://www.icsu.org/publications/open-data-in-a-big-data-world

Schmidt, S. (2009) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90-100.

Shavit, A and Ellison, AM (2017) Towards a Taxonomy of Scientific Replication in Shavit, A and Ellison, AM (eds) *Stepping into the Same River Twice: Replication in Biological Research*. Yale University Press.

Simmons, J.P., Nelson, L. D. and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 (11): 1359–1366.

The Economist (2013) Trouble at the Lab. Last Accessed 18 Jauary 2018, URL: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Turner, Robert, and Daniel De Haan (2017) Bridging the Gap between System and Cell: The Role of Ultra-High Field MRI in Human Neuroscience. In *Progress in Brain Research*, edited by Tara Mahfoud, Sam McLean, and Nikolas Rose, 233:179–220. Vital Models. Elsevier. http://www.sciencedirect.com/science/article/pii/S0079612317300493.

Zahle, J. (2018) Values and Data Collection in Social Research. *Philosophy of Science* 85 (1): 144-163.