# Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing

**Bart Penders [1,*]** [ID], **J. Britt Holbrook [2]** [ID] and **Sarah de Rijcke [3]**

1   Department of Health, Ethics & Society, School of Primary Care and Public Health (CAPHRI), Maastricht University, PO Box 616, NL-6200MD, Maastricht, The Netherlands
2   Department of Humanities, New Jersey Institute of Technology, Newark, NJ 07102, USA
3   Centre for Science and Technology Studies (CWTS), Leiden University, PO Box 905, NL-2300AX Leiden, The Netherlands
*   Correspondence: b.penders@maastrichtuniversity.nl; Tel.: +31-(0)43-388-2457

check for updates

**Abstract:** The increasing pursuit of replicable research and actual replication of research is a political project that articulates a very specific technology of accountability for science. This project was initiated in response to concerns about the openness and trustworthiness of science. Though applicable and valuable in many fields, here we argue that this value cannot be extended everywhere, since the epistemic content of fields, as well as their accountability infrastructures, differ. Furthermore, we argue that there are limits to replicability across all fields; but in some fields, including parts of the humanities, these limits severely undermine the value of replication to account for the value of research.

**Keywords:** Replicability; replication; reproducibility; reproduction; epistemic pluralism; accountability; humanities

## 1. Introduction

Although it would be possible to suggest that replication is simply an epistemic matter, the increasing drive for replication is also a political project and should be viewed in connection with the pursuit of open science and scholarship. Open scholarship, in its broadest incarnation [1], seeks to open up both knowledge-making and knowledge-communication and -translation processes to peer and public scrutiny and participation. Researchers pursue open scholarship (and science) for many reasons, including but not limited to: (1) the ambition to democratise knowledge as a process and as a product to allow as much expertise and as many types of expertise as possible to contribute; (2) the desire to reveal dubious activities or decisions under the heading of fraud, misconduct or questionable research practices; (3) the attempt to explore and expose a large variety of biases or other prejudices—conscious or otherwise; and (4) the effort to shift capital, ownership and stewardship of science, both content and containers, back into the public realm. Policy makers aim to encourage open science because openness is supposed to contribute to the trustworthiness of science and scholarship among peers and publics alike by: (1) avoiding too narrow, overly reductionist or simply incomplete work; (2) exposing and reducing the largely intentionally fraudulent or invalid and the largely unintentionally invalid; and (3) replacing undesirable accountability and evaluation structures with those that steer researchers toward further openness.

When we speak of 'openness' or 'open science' or 'open scholarship' generally, then, we are referring to a family of activities and policies oriented around the values of democratising and legitimising knowledge production and use. It is a phenomenon that includes but extends beyond the Open Access (OA) movement. Conceived broadly, various approaches to citizen science, crowdsourcing,

responsible research and innovation, citizen consultations, and the like should be included under its rubric. The concern to get the science right, a prominent feature of the drive for replication, is motivated not only by the pursuit of truth for its own sake, but also by the potential impacts on both science and society of getting the science wrong. If most published research findings are false, then policies based on them are likely to be mistakes. Even if most published research findings are in fact true, and so valuable for informing policy and decision making, if the public believes they are false, it undermines the legitimacy of scientific research and the policies informed by it. Distrust in science would undermine science itself. Thus, the same problems increased openness seeks to solve have also contributed significantly to the current replication drive and its framing as a 'crisis'.

Whether or not the label 'crisis' is legitimate is a reasonable question and qualifications differ remarkably [2–7]. However, there obviously is momentum generated through the rise of the reproducibility movement and the new socio-technical (or methodological) infrastructure that is emerging in science to avoid irreplicability. This infrastructure includes the use of the preregistration of studies, new demands on communicating methodological details, and efforts to open up data for inspection as well as for re-use. Replication concerns emerged first in psychology but rapidly spread to other fields, most notably the life sciences. In both areas, as well as far beyond them, the conversation on open science and replicability partially merged in the sense that increased openness was generally considered to alleviate replication-related concerns. Open science, as a practice and an idea, is rapidly going mainstream. It has been written into a number of policy documents, including discussions of the European Commission's Horizon Europe, Plan S to promote open access, and the newest version of the Dutch Code of Conduct on Research Integrity, which clearly positions openness as the norm across all research.

Openness ambitions are growing, and policy makers are rolling out new policies to encourage openness across entire academic communities, regardless of discipline, epistemology, object of study, digital and financial capabilities and more. As a consequence, expectations about how to handle, evaluate or perceive replication are travelling as well. In this article, we argue that openness does not guarantee replication. Further, building from our earlier work [8,9], we argue that replication and replicability are not always possible, regardless of how much openness one can muster. Beyond the purely epistemic question of whether replication is possible across different fields of research, we also argue that the replication drive qua policy for research is especially questionable in the hermeneutical social sciences and humanities. We focus on these fields because they have received little to no attention in the replication debate before, yet they have been included quite explicitly in recent calls to extend replicatory expectations. We have an explicitly ethico-political concern here. If fields of research exist for which replication is an unreasonable epistemic expectation, then policies for research that universalise the replication drive will perpetrate (some might say perpetuate) an epistemic injustice, ghettoising the humanities and hermeneutic social sciences as either inferior research or not really research at all.

In the sections below, we first clarify the replication terminology we are using and how it relates to how others are writing about and have written about replication. We then discuss the epistemic limits of the possibility of replication and replicability in the humanities. Subsequently, we address whether replication and replicability policies are desirable for the humanities. Finally, we discuss how accountability and quality in research can be conceptualised given limits on replication and replicability in the humanities.

In essence, we argue that replication talk, replicatory expectations, and replication and replicability policies should be contextualised in relation to specific epistemic communities, rather than applied across the board, as if there were a single approach to knowledge and truth. This more context-sensitive appreciation of replication may offer a better understanding of how we should understand and mobilise openness. After all, open scholarship, open science and replication are not only epistemic, but also ethico-political projects, pursuing, in their broadest incarnation, a radical reform of the organisation of

science and scholarship in general and the ways in which researchers give account of what they do and how they can be held accountable.

## 2. Replication Talk

Although terms such as 'replication', 'reproduction', 'robustness', and 'generalisability' are common terminology in discussions of scientific quality control, they are not defined consistently and sometimes not defined at all [10,11]. For instance, reproducibility is often confused or conflated with replicability, and even these two concepts do not cover all possibilities. Some may suggest that terminological wrangling is beside the point. However, particularly when it comes to making policies for research, the language we use matters. This is especially true if, as we argue, various epistemologies are supposed to fall under umbrella policies for research. For these reasons, we attempt to bring some order to replication talk.

The US National Science Foundation (NSF) defines reproduction as "the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator," while replication refers to "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected" [12]. According to NSF, reproducibility does not require new data, while replication does. Of course, replication can occur under the exact same circumstances, or data can be samples from a similar, rather than the same population. In the latter case, provided the replication succeeds, robustness and generalisability are built [13].

While this may seem clear and straightforward, multiple conflicting definitions coexist. For instance, Claerbout and Karrenbach [14] define reproduction as the act of "running the same software on the same input data and obtaining the same results" and argue that it may even be automated (see also [15]). Rougier defines replication as "running new software [...] and obtaining results that are similar enough" [15]. The difference between reproduction and replication here lies in the code used, not the data.

Crook et al. [16] follow the Association for Computing Machinery (ACM, see also [17]) and discriminate between reproducibility, replicability and repeatability. To them, repeatability means that researchers can reliably repeat their own computation and obtain the same result. Replicability entails that another independent researcher or group can obtain the same result using the original author's own data. Reproducibility entails that another independent researcher or group can obtain the same result using data or measurements they conduct independently.

Peels and Bouter [18] also briefly define two terms: replicability and replication. They define replicability as a characteristic of the communication of an approach, and replication is the act of repeating that approach. On this account, replicability ensures the conditions for the possibility of replication. Peels and Bouter's use of the term replication differs significantly from Bollen et al.'s, Clearbout et al.'s and Crook et al.'s in the sense that all others include the requirement or ability to obtain the same results (even if they choose to define the process as reproducibility, in the case of Claerbout et al. and Crook et al.). Peels and Bouter separate ability, act and result. The ability, replicability, deals with how approaches and methods are communicated. The act, replication, can take place in two specific ways: the direct replication (replication using the same study protocol, but collecting new data), and conceptual replication (replication using a modified, or different protocol, to collect new data). While their definitions are clear, this creates confusion in the sense that it conflicts with other uses of the same terminology. The statement "This approach is replicable" means different things to Peels and Bouter, Bollen et al., Claerbout et al., Crook et al. and presumably many others. We summarise this confusion in Table 1.

**Table 1.** Conflicting definitions of replicability.

| Author | Statement | Meaning |
| --- | --- | --- |
| Bollen et al. | This approach is replicable | This approach is described so that others can repeat it using new data and obtain the same (or similar enough) result. |
| Claerbout et al. | This approach is replicable | This approach is described so that others can repeat it while using new tools to obtain the same (or similar enough) result. |
| Crook et al. | This approach is replicable | This approach is described so that others can repeat it using the data produced by the original authors and yields the same (or similar enough) result. |
| Peels and Bouter | This approach is replicable | This approach is described so that others can repeat it. |

To complicate things further, Goodman et al. [10] do not speak of replicability at all, yet they distinguish between three different types (or levels) of reproducibility: methods reproducibility, results reproducibility and inferential reproducibility. The first overlaps with Peels and Bouter's definition of replicability, the second with Bollen et al.'s replicability and the third refers to the ability to draw the same conclusion, regardless of whether the data or results overlapped. Regardless of the definition used, it is important to note that neither 'replicable' nor 'reproducible' is a synonym for 'true', because, as Leek and Peng [19], Penders and Janssens [20] and Devezer et al. [21][1] argue replicable, and even replicated, research can be wrong.

Since Peels and Bouter are willing to attribute the label 'replicable' to studies or bodies of work, while remaining agnostic about the outcome of the act of replication, their conceptualisation of replicability sets the lowest bar. Additionally, in contrast to all other literature discussed in this section, they explicitly mention that they devised their definitions to apply to humanities research as well. Finally, Peels and Bouter seem to be giving voice to a wider trend in the governance and accountability of research in general, a trend that ties discussions on rigour, replicability and quality with a certain degree of uniformity, even, for instance, into the rationales supporting funding programmes and policies. We will return to this policy trend in the final section of the paper. Accordingly, for the remainder of this text, we will use the terms replicability and replication in line with Peels and Bouter. The key characteristic of a replication study according to Peels and Bouter is that it attempts to answer the same question as the initial study, with replicability as a requirement for initiation [18,22–24].

## 3. Possibility

Are replicability and replication characteristics of scientific and scholarly reporting and work that all studies, regardless of discipline, origin, topic or style can and should achieve? Here, we will first discuss the possibility of replicability and replication,[2] drawing from a few landmark studies in the sociology and history of science, with questions of desirability reserved for the next section.

In his landmark study of replication, sociologist of science Harry Collins studied and documented how scientists replicate, which standards they establish to do so, and which challenges they encounter when pursuing replication and replicability of their work [25]. Collins assesses the value of replicability to scientists as follows: "Replicability [...] is the Supreme Court of the scientific system" (p. 19), yet qualifies the practice as more complex, less straightforward and inherently social. Collins conceptualises the practice of doing replication as one in which it matters who repeats an experiment or study (who counts as a legitimate peer, and how do we evaluate his/her relationship to the original experimenter?),

---

[1] Devezer et al. phrase this as follows: "reproducibility does not imply convergence to the truth (nor vice versa) and hence we must be careful not to substitute one for the other" (p. 22).
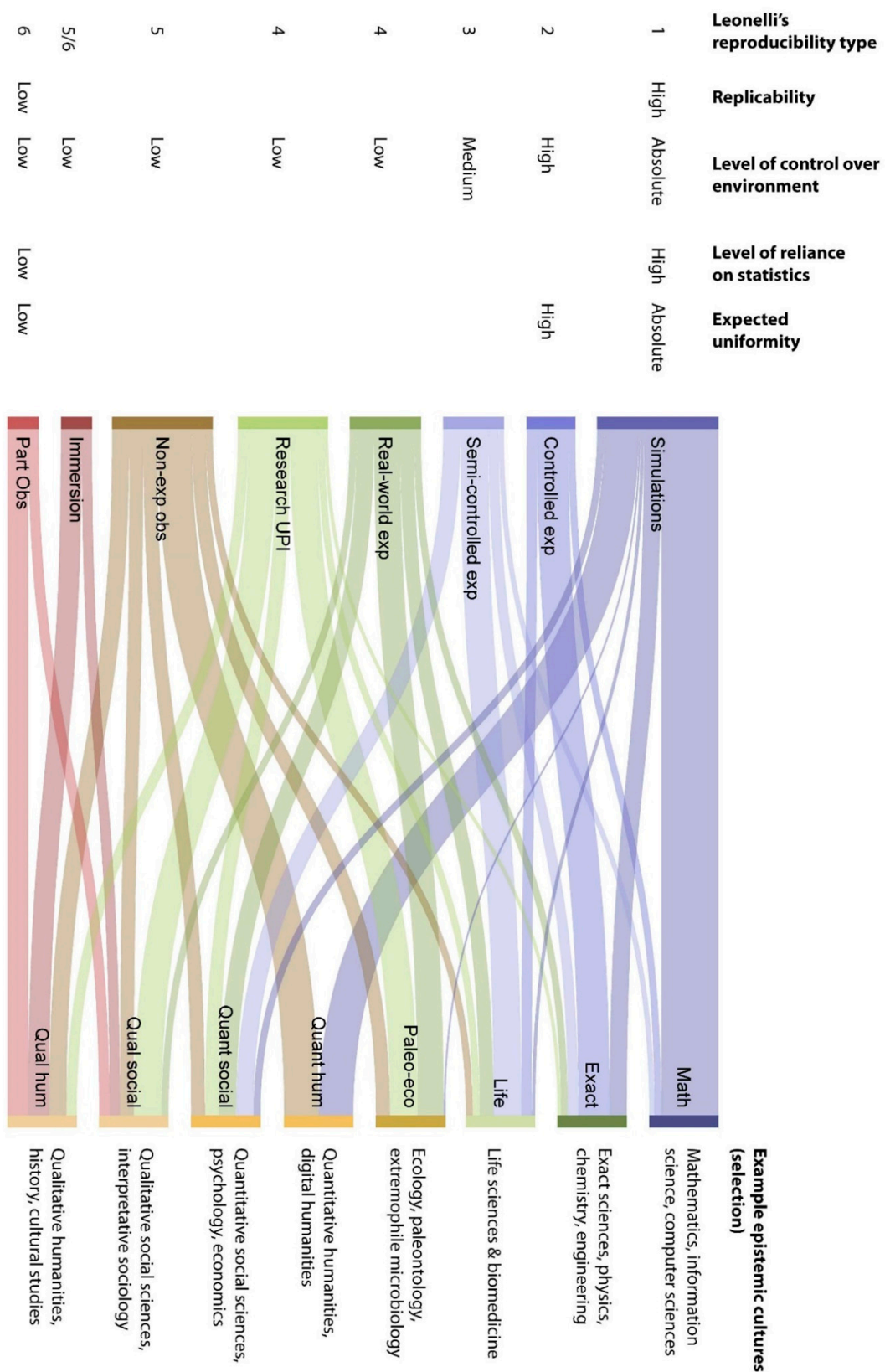[2] With the important caveat that, in this paper, we limit ourselves to empirical studies only.

which tools were used to do so (what is a credible tool and what makes a tool credible?), what can be written down in a study protocol and what cannot (if the outcomes of an experimental setup rely on unknown unknowns, how can they be incorporated in the replication?), what can be communicated or taught in practice (which parts of required knowledge and expertise are formal, and which one informal or tacit, limiting their formal exchange?) and what to do when experts disagree about whether a replication constitutes a replication or when both positive and negative (successful and failed) replications co-exist.

Collins documents (pp. 51–78) how British scientists are unable to build a laser from American written instructions (as listed in published papers, manuals, blueprints and much more). This was neither an issue of different conventions across the Atlantic nor a careful scheme to prevent one group of researchers from attempting replication. The Americans were, in fact, helping the British to make this happen and providing all the information they could, answering every possible question truthfully. Collins notes that the only situation in which a replicated laser was made to work was through direct personal contact between one of the original researchers and the ones attempting the replication. Attempts to 'make it work' via the use of a middleman all failed. Something is made to transfer face-to-face, through familiarity and proximity, which is not transferred through immensely detailed instructions or data sets. Drawing on Polanyi, Collins calls this tacit knowledge, which he describes as "our ability to perform skills without being able to articulate how we do them" (p. 56). Not knowing how to articulate how we do something limits our ability to explain how we do it, provided we are aware of it. A lot of our tacit knowledge is knowledge of which we are not aware that we have and, when made aware, are unable to articulate. Replicating the construction of a laser—physics and engineering, two fields considered ultimately universal and replicable—required the transfer of tacit knowledge.

The problem of tacit knowledge emerged at the dawn of experimental science in the 17th Century, where similar issues plagued air-pumps (vacuum pumps): "The pumps that did work, could only be made to operate with a group which already had a working pump. Replication, which is often held to be the touchstone of the experimental form of life, turned out to be very messy indeed" [26] (see also [27]). Tacit knowledge cannot travel as text, but has to be exchanged through apprenticeships, supervised practice and "is invisible in its passage and those who possess it" (p. 74). The exchange of tacit knowledge requires a relationship, yet that relationship may influence how others evaluate the independence and credibility of the subsequent replication.

Collins thus demonstrates practical limits to replicability: not everything can be documented so that others can repeat it. Without reaching consensus on what the 'right' result is, no consensus on the 'right' approach can be reached (see also [28]). That is not an act of deception, but rather a limit to how knowledge, expertise and skill can travel. Tacit knowledge extends into every epistemic domain including the humanities, from theoretical physics to ethnography, but the ratio of formal versus tacit knowledge will differ significantly between epistemic domains and consequently between types of replication attempts.

Rather than asking what can be achieved in practice, as Collins did, Leonelli [29] asks where replication can be achieved and whether limits on replication (including those proposed by Collins) differ between epistemic cultures. Leonelli uses reproducibility terminology, but her argument can be extended into the vocabulary used here as well. In her critical discussion of the limits of replicability as a potential criterion for the quality of research, she distinguishes at least six ways of doing empirical research [29]. These are: (1) software development, (2) standardised experiment, (3) semi-standardised experiments, (4) non-standard experiments or research based on rare, unique, perishable or inaccessible materials, (5) non-experimental case descriptions and (6) participant observation (also see Figure 1). Replication and replicability, she argues, are first and foremost completely different entities with different characteristics, criteria and requirements in all six types of work. Secondly, replication and replicability carry completely different weights in all six. Leonelli thus proposes a situated understanding and a situated valuation of replication and replicability.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Leonelli's reproducibility type | 6 | 5/6 | 5 | 4 | 4 | 3 | 2 | 1 |
| Replicability | | Low | | | | | High | High |
| Level of control over environment | Low | Low | Low | Low | Low | Medium | Absolute | Absolute |
| Level of reliance on statistics | Low | | | | | | High | |
| Expected uniformity | Low | | | | | High | | Absolute |



**Figure 1.** A taxonomy of replicability. This taxonomy is a visual representation of the ways in which nearly all disciplines, fields or epistemic cultures differ in the ways in which they make knowledge and thereby in their capacity to replicate successfully and or attain replicability. The epistemic cultures connected to replicability types are listed in no meaningful order.

Like Collins, Leonelli is after a systematic understanding of replication across all ways of knowing. Unlike Collins, at least in his early work on replication, Leonelli sees and discusses a number of research approaches that are more dominant in the humanities than, for instance, the life sciences. Although Leonelli does not explicitly make this connection, we suggest that hermeneutical social science and humanities research would populate the categories "non-standard experiments & research based on rare, unique, perishable, inaccessible materials" (e.g., history, studies of public opinion or morality), "non-experimental case description" (e.g., history, arts, philosophy, interpretative sociology) and "participant observation" (e.g., interpretative sociology, anthropology).

These three categories exhibit a series of characteristics that distinguish them from the former three. First, in the context of simulation and (semi-)standardised experiment, researchers have substantial control over the environment in which knowledge making takes place. In a simulation, that control is absolute, and in controlled experimental setups it really is not, but control is still high. In stark contrast, non-standard experiments, non-experimental setups and participant observation offer little to no control over the environment. Second, Leonelli demonstrates, statistics are of primary importance as an inferential tool in the first three categories, but of diminishing importance in the last three. Of course, we must add, the lines between the categories are fuzzy, and combined they represent a gradient of control, rather than six distinct levels. However, as a heuristic to understand the situated limits on replicability and replication, Leonelli's distinction is quite valuable.

In non-standard experiments and non-experimental systems, replicability exists as a theoretical possibility. However, because of the lack of control researchers have over the environment in which research takes place, actual replication is contingent on circumstances beyond their control. The outcome of that replication study thus cannot confirm or refute the original study. In the last category, "participant observation", Leonelli argues that "different observers are assumed to have different viewpoints and produce different data and interpretations," (pp. 12–13) which means that replicability cannot be reached (and thereby replication cannot be attempted).

In Collins' recent autobiographical reflection on his original study of replication [30], he argues that interpretative sociological work may also be theoretically replicable (provided one immerses oneself in the same or very similar [sub-]culture at the same time, given that all cultures change); but given such minimal control over studied social groups and cultures, an actual replication will be quite unlikely. In this sense, Collins' immersion is epistemically distinct from Leonelli's participant observation, in terms of replicability potential—even if only in theory. Thus, where Collins proposes universal limits on replicability in every knowledge domain, both Collins and Leonelli argue that some fields display characteristics that decrease the value replicability has for asserting the status of its knowledge claims. Among those fields are some of the humanities and social sciences.

Leonelli initiated a taxonomy by presenting her typology of replicability, and we expand upon it here. We drew apart her category of non-standard experiments or research based on rare, unique, perishable or inaccessible materials into two: "non-standard experiments" and "research based on rare, unique, perishable or inaccessible materials", since they present different types of limitation on replication and replicability. We added "immersion", drawn from Collins [30,31] to allow for a distinction in replicatory ambitions in qualitative fieldwork with Leonelli's participant observation [29]. Of course, epistemic cultures grow, change and evolve; they diverge, converge, merge or spawn new approaches. Epistemic cultures can be delegitimised and legitimised (e.g., the delegitimation of phrenology and the legitimation of gender studies). Distinctions of entities, as in other taxonomies, are in part, arbitrary and require consensus. Therefore, any taxonomy of replicability is difficult, situated, contestable and controversial—as it should be. The same goes for any visualisation of that taxonomy. In the alluvial diagram (see Figure 1), the taxonomy of replicability is connected to a diverse set of fields and disciplines. None of the connections between replicability types and epistemic cultures are exclusive, meaning that each replication type hosts examples from multiple epistemic cultures and, that in all epistemic cultures, many different replicability types can be identified.

This taxonomy does not distinguish between exploratory and confirmatory research. In every epistemic culture, both types of research coexist and feed into one another. Where exploratory research approaches data without clear questions in mind, generating knowledge and theory (sometimes also called hypothesis-generating), confirmatory research seeks to test theories or refute concrete hypotheses. The ratio between the two types of research is not uniform across cultures, nor are the tools to test theories or refute hypothesis. Importantly, not all qualitative research is exploratory, even if decisions on whether a hypothesis or theory is refuted look differently [32].

Similar to Collins [30], Lamont and White also report that some, but not all, fields that employ qualitative research methods "attend to the issue of replicability" [32] (p. 14), although they do not distinguish between the epistemology informing the method (they identify a contrast between constructivist and positivist epistemologies in qualitative research). Their definition, in line with the ones in Table 1, would be *This approach is described so that others can check it and critically (re)examine the evidence*. In their call to make qualitative research more transparent, Elman and Kapiszewski do refer to researcher epistemologies as relevant differences, including the caveat that methodological and data transparency has limited value in the context of interpretative research:

> "[I]nterpretive scholars argue that this step is unachievable, because the social world and the particulars of how they interact with it are only accessible through their exclusive individual engagement and therefore cannot be fully captured and shared in a meaningful way." [33]

Political scientist Gerring [34] adds that expectations of low uniformity (see also [31]) reasonably legitimise the use of qualitative methods. Low uniformity makes comparison difficult, leading him to conclude that under such conditions, the stability and thus replicability of research is low (p. 114). He mentions historical, legal and journalistic accounts explicitly, as well as single-case studies. In the same report (on quality, standards and rigour), Comaroff [35] dives into social anthropology to conclude that ethnographers and participant-observers can and should devote a lot more attention to explicate their procedures fully, not because it would create replicability but because it would allow scrutiny. This is in line with Wythoff [36], who writes

> "[M]ethod is interesting in the humanities not because it makes possible replicability and corroboration as it does in the sciences, but because it allows us to produce useful portraits of the work we do: our assumptions, our tools, and the assumptions behind our tools." (p. 295)

Similarly, and in disagreement with Collins, Strübing argues that replication is valuable where it can be achieved, but that this is not the case in qualitative interpretive research [37]. Requiring it in the form of a replication drive would amount to an active devaluation of qualitative interpretative research [38].

## 4. Desirability

Having established limits to replication and replicability, we can ask when and where it would be a sensible policy to pursue or demand it and what such policies ask of research practices. Calls to extend existing norms for replicability and replication into new fields are inherently normative. They invoke images of good research and proper, accountable and correct, science, where replication movements are the preferred technique to demonstrate this correctness and display accountability. For instance, Peels and Bouter [18] offer the following argument in response to the question of whether replication is desirable in the humanities:

> "Yes. Attempts at replication in the humanities, like elsewhere, can show that the original study cannot be successfully replicated in the first place, filter out faulty reasoning or misguided interpretations, draw attention to unnoticed crucial differences in study methods, bring new or forgotten old evidence to mind, provide new background knowledge, and detect the use of flawed research methods. Thus, successful replication in the humanities also makes it more likely that the original study results are correct." (p. 2)

We suggest, however, that this overstates the case. Successful replication of a study in the humanities would increase our confidence in the reliability of the results—that is, in our ability to replicate the study. However, the reliability of a study does not make it more likely that the results of the original study are correct (true). Peels and Bouter do admit that a study and its replication could both be wrong in the sense that they could be equally biased, but they neglect that both could be wrong in the sense of yielding results that are simply false.

The value of a successful replication (with a result that overlaps sufficiently with the original) or a failed replication (with a result that differs from the original) in the humanities or hermeneutical social sciences is not universal. In all cases, it is important to note that a failed replication does not constitute bad science, a situation shared by all disciplines. Where Leonelli argued that replicability does not constitute a valid quality criterion for science and can be considered variable across epistemic cultures, Leung argued that for qualitative research, equating reliability and replicability is "challenging and epistemologically counter-intuitive" [39].

To seek replication anyway, ignoring all limitations discussed above, constitutes a forced unification of epistemologies and a forced unification of the values they contain—including but not limited to the value of replicability. This forced unification is in fact, well-described and opposed [40]. Forcing notions of replicability onto epistemic communities that do not share them, including but not limited to the humanities, imposes a form of monism that denies the possibility for alternative forms of knowledge making or interpretation. To Chang, the sciences and humanities serve the goal of giving us accounts of the world that serve whatever aim we have. The monistic account should not be the aim itself. Rather, he argues, "science can be served better in general by cultivating multiple interactive accounts [40] (p. 260).

Replicability in itself is not a virtue. Leonelli [29] writes that "[r]esearchers working with highly idiosyncratic, situated findings are well-aware that they cannot rely on reproducibility as an epistemic criterion for data quality and validity." (p. 11). In the humanities, and especially in the interpretative or constructivist epistemic cultures it hosts, a lot of research can be recognised by the contextual character of knowing. In many such contexts, the prior delineation of study populations, hypotheses, and data-extraction protocols would defeat the very purpose of the research [41]. In addition, research value is often generated by adding to the diversity of arguments. Of course, in many epistemic cultures, replicability is of uncontested value, and, under some circumstances, replicability would be useful in some humanities too (for a selective list of examples, see [18,22–24] as well as the budding taxonomy of replicability in Figure 1). But to require it of others would be harmful. Understanding cultural phenomena, such as migration or security, depends on the diversity of arguments and positions to help develop global solutions. Interpreting classical or medieval literature requires the continuous development of alternative, competing and often incommensurable readings, and interpreting the writing of philosophers, cultural movements and trends similarly benefit from the diversity produced. The humanities, and—in Collins' words—the humanities' approach to social science, pursue not stability, but "persuasive novelty" [31].

One step further, constructivist epistemologies are found mostly in the social sciences and humanities. They propose that knowledge is not simply found or discovered, but instead it is constructed (built) by researchers by drawing together material and social elements to support knowledge claims. It exists in a large number of variations, yet its focus on the situated character of knowing can be recognised across all of them. Observing a work of art (whether looking at a Rembrandt, reading Shakespeare, listening to Tchaikovsky, or smelling a Tolaas) is a situated act that only acquires meaning in the context of the relevant circumstances. Attending to the situated character of the act also matters for humanities research that relies on such observations, whether directed at pre-modern French poetry or organic modernism in Danish design. In the context of social science, interaction between individuals, between individuals and objects and dynamics in groups and cultures similarly limit intersubjectivity.

Ethnographic research, interpretative sociology, and many more highly empirical research approaches do not pursue replication and are, in the definition by Peels and Bouter (as well as all others) irreplicable. That does not mean that the products of this type of research are not subject to critical scrutiny or do not need to live up to expectations of quality and rigour. Rather, the pursuit of rigour is cast in terms of accountability, rather than replicability. Participant observers (to use Leonelli's term) are required to describe to the best of their abilities, the situation in which observations took place in order to allow readers to "evaluate how the focus of the attention, emotional state and existing commitments of the researchers at the time of the investigation may have affected their results" [29].

In this context, calls for the preregistration of studies and similar strategies to curtail bias [42] lose relevance. Where neither objectivity, nor intersubjectivity is pursued, the goals of preregistration fall away. Hence, just as the possibility of replication and replicability should be decided relative to the research approach employed, so too the desirability of replication and replicability should be situated and judged relative to different fields.

## 5. Politics of Accountability

There is nothing wrong with being concerned about the state of science and scholarship in 2019. Indeed, calls for replication drives or to promote open science are often highly commendable (cf. [42]). We assume that pleas to extend accountability regimes such as replication [18,22–24] and preregistration [43] into new epistemic cultures are motivated by such genuine concern. However, these epistemic extensions carry with them a moral imperative: scientists' responsibility to ensure the accuracy of their collective's claims through replication. While shared by many, this particular incarnation of this moral imperative is not without its opponents (see e.g., [29]).

There is a shared responsibility by all stakeholders to ensure rigour and high-quality reporting of research. This goes for all the sciences and humanities. However, replicability—even according to the most lenient definition by Peels and Bouter—is not attainable for a large portion of the epistemic cultures that together make up the humanities and social sciences. Diligent description of methods, data gathering and analysis, and the circumstances under which all of them took place—with a keen eye for reflexivity—all add to the rigour of reporting research and the potential reliability of that research. But they do not and need not automatically generate the ability to replicate a study.

Uncritically extending replication as a moral imperative across all epistemic domains risks damaging the plural epistemologies that the sciences and humanities contain and the value that plurality represents [40]. Peels and Bouter advocate for a replication drive in the humanities, calling it an "urgent need." According to them, funding agencies should demand that any primary studies they fund in the humanities are replicable and begin funding replication studies; journals should publish replication studies, regardless of results; and humanistic scholars and their professional organisations should "get their act together" [18]. Nosek et al. [42] do not explicitly seek to impose a replication drive on the humanities; however, they lament the fact that there is no central authority that can impose a replication mandate on science:

> "Unfortunately, there is no centralized means of aligning individual and communal incentives via universal scientific policies and procedures. Universities, granting agencies, and publishers each create different incentives for researchers. With all of this complexity, nudging scientific practices toward greater openness requires complementary and coordinated efforts from all stakeholders." (p. 1423)

Even if we grant that Nosek et al. intend to suggest only to "nudge" all science (excluding the humanities) toward openness and replicability, the fact that they contrast their suggestion with the current "unfortunate" lack of a central means to align all policy is telling. Their rhetoric brings our reservations about a universal replication drive into stark relief: even if we recognise that different areas of research have—and should have—different ways of knowing, it remains tempting to think that what is good for one area of research must be good for all research. We think this temptation

should be resisted. This is especially important when it comes to developing policies for science and scholarship. Given the current status of the humanities relative to other scientific and technical fields, resisting policies that may further marginalise the humanities is particularly vital. From the fact that a small portion of research in the humanities may be replicable, it does not follow that all research in the humanities ought to be replicable. To adopt policies that require replicability of all funded research would rule out funding for the vast majority of research in the humanities. As Collins [25] already demonstrated, replicability does not guarantee correctness, to which Leonelli added that replicability does not align with quality [29]. In a similar vein, Derksen and Rietzschel [28] argued that replication cannot and should not serve as control technologies since they cannot serve as "tests of whether an effect is 'true' or not—or, worse, if non-replications are taken to mean that 'something is wrong'" (p. 296). All researchers, including those in the humanities, should be able to account for all elements in their research design, and they should understand its consequences. But the crucial point is that many humanities approaches (including their practices of reporting) allow researchers to deal with the (im)possibility of replication by giving particular accounts of the consequences of methodological decisions and the role of the researcher.

Policies suggesting, requiring or demanding replicability, or even successful replication (or plain 'replication' in the definition by Bollen, see Section 1), impose monism [40]. They inadvertently, or purposefully, claim that one set of epistemic cultures is superior to another, delegitimising the latter. In contrast, more balanced positions with respect to replicability respect such pluralism and position replicability merely as one of many values meant to support the quality, rigour and relevance of research. Compare the following recent statements from various funding agencies. The Netherlands Organization for Scientific Research (NWO) claims:

> "Not all humanities research is suitable for replication. NWO is aware of the discussion currently taking place about this within the humanities and expresses no preference or opinion about the value of various methods of research. Where possible it wants to encourage and facilitate the replication of humanities research: this should certainly be possible in the empirical humanities."[3]

NWO claims awareness of the ongoing discussion of replication within the humanities, claims neutrality, then professes its desire to encourage replication of humanities research. The German Research Foundation (DFG), on the other hand, while suggesting that replication may be possible within the humanities, is careful to say that replication is not the only criterion for good research. DFG, in fact, explicitly mentions "theoretical-conceptual discussion and critique", techniques commonly used in the humanities, as viable test procedures [44]. DFG also explicitly claims, "Replicability is not a general criterion of scientific knowledge".[4] The US National Institutes of Health explicitly confine their plans to enhance reproducibility to biomedical research [45]. The US National Science Foundation (NSF) has issued a series of Dear Colleague Letters that encourage replication in specific subject areas.[5] The recently released National Academies report on "Reproducibility and Replicability in Science" [46] is striking for a number of reasons. First, the report limits reproducibility to computational reproducibility. Second, it suggests that replicability in science is an inherently complex issue. Third, the word 'humanities' appears nowhere in the report. Finally, the report itself is an exercise in the philosophy of science (answering questions such as "What is Science?").

Humanities and interpretive sociological research are different from the sciences not because of some sort of secret sauce, but because the objects of study, and the questions asked, often, but not always, do not allow replication or even replicability. As a consequence, humanities research needs to

---

3    https://www.nwo.nl/en/news-and-events/news/2019/03/third-round-in-pilot-replication-studies-now-includees-the-humanities.html (accessed on 4 June 2019)
4    Translated from German "Replizierbarkeit ist kein generelles Kriterium wissenschaftlicher Erkenntnis".
5    For an example, see https://www.nsf.gov/pubs/2018/nsf18053/nsf18053.jsp?org=SBE (accessed on 4 June 2019).

be organised differently to still be able to give account and be held accountable. Leonelli hints at this in her critique of replication [29], that when researchers realise that replication cannot be an indicator of quality, they:

> "instead devote care and critical thinking on documenting data production processes, examining the variation among their materials and environmental conditions, and strategize about data preservation and dissemination. Within qualitative research traditions, explicitly side-stepping [replicability] has helped researchers to improve the reliability and accountability of their research practices and data". (p. 14)

Drawing from this tradition, Irwin has, in an attempt to respect epistemic pluralities, suggested quality criteria beyond replication that span many of the social sciences and humanities and include responsibility, public value, cognitive justice and public engagement [47]. They can be assessed only in dialogue. Even though replicability cannot be claimed to be possible or desirable across the whole of the sciences and humanities, detailed scrutiny of existing principles and practices for assessing proper conduct in the humanities is needed for the development of sensible, yet plural, methods for quality assessment. In this plural accountability toolbox, there is plenty of room for replicability and replication, including a lot of infrastructure that accompanies it under the wider umbrella of Open Science. The accountability toolbox is, however, a lot bigger, and humanities scholars have been diligently filling it with diverse tools suited to diverse research practices. Replicability and openness and their characteristics, limitations and status as a moral imperative must be made to fit research practices to allow them to contribute locally, rather than the other way around. Not allowing humanities scholars to engage in this process, forcing the humanities communities to adopt epistemic standards drawn from the sciences, disqualifies humanities and other scholars as legitimate knowing subjects in their own field: an epistemic injustice [48]. This does not mean that the pursuit or support of replication as a technology of accountability is unjust. It means that denying the epistemic authority of the humanities is unjust, even if one disagrees with how they organise and reward knowledge making.

If we need to limit our urge to impose a universal moral imperative of replication or openness on all of science, then we need even more to limit our urge to impose universal policies that fail to account for local (epistemic) differences. It is one thing for researchers to virtue signal openness in exchanges on Twitter. It is something decidedly different to advocate for policies that would force subgroups of researchers to abandon the disciplinary standards of their fields. As researchers, we have different obligations to different stakeholders. We have obligations to our departments and to our universities. We have obligations to our funding agencies. We have obligations to society. We also have obligations to our fields of study. It is difficult to imagine researchers who fail to recognise all of these obligations. We should, indeed, expect all researchers to be able to account for their research, including its societal value. We should not, however, appeal to a nebulous notion of an imagined unified, monistic science in order to justify policies that elide justifiable epistemic differences among various fields of research.

## References

1. Neylon, C. Openness in scholarship: A return to core values? In *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices: Proceedings of the 21st International Conference on Electronic Publishing*; Chan, L., Loizides, F., Eds.; IOS Press: Amsterdam, The Netherlands, 2017; pp. 6–17. [CrossRef]
2. Lash, T.L.; Collin, L.J.; Dyke, M.E. The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Curr. Epidemiol. Rep.* **2018**, *5*, 175–183. [CrossRef]
3. Randall, D.; Welser, C. *The Irreproducibility Crisis of Modern Science: Causes, Consequences, and the Road to Reform*; National Association of Scholars: New York, NY, USA, 2018.
4. Harris, R. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*; Basic Books: Hachette, UK, 2017.
5. Saltelli, A.; Funtowicz, S. What is science's crisis really about? *Futures* **2017**, *91*, 5–11. [CrossRef]
6. Fanelli, D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2628–2631. [CrossRef] [PubMed]
7. Redish, A.D.; Kummerfeld, E.; Morris, R.L.; Love, A.C. Opinion: Reproducibility failures are essential to scientific inquiry. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 5042–5046. [CrossRef] [PubMed]
8. Holbrook, J.B.; Penders, B.; Rijcke, S. The humanities do not need a replication drive. *CWTS Blog*. 2019. Available online: https://www.cwts.nl/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive (accessed on 21 March 2019). [CrossRef]
9. Rijcke, S.; Penders, B. Resist calls for replicability in the humanities. *Nature* **2018**, *560*, 29. [CrossRef] [PubMed]
10. Goodman, S.N.; Fanelli, D.; Ioannidis, J.P.A. What does research reproducibility mean? *Sci. Trans. Med.* **2016**, *8*, 312–341. [CrossRef]
11. Kenett, R.S.; Shmueli, G. Clarifying the terminology that describes scientific reproducibility. *Nat. Methods* **2015**, *12*, 699. [CrossRef]
12. Bollen, K.; Cacioppo, J.T.; Kaplan, R.M.; Krosnick, J.A.; Olds, J.L. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*; Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences; National Science Foundation: Arlington, VA, USA, 2015.
13. Clemens, M.A. The meaning of failed replications: A review and proposal. *J. Econ. Surv.* **2017**, *31*, 326–342. [CrossRef]
14. Claerbout, J.F.; Karrenbach, M. Electronic documents give reproducible research a new meaning. *SEG Expand. Abstr.* **1992**, *11*, 601–604. [CrossRef]
15. Rougier, N.P.; Hinsen, K.; Alexandre, F.; Arildsen, T.; Barba, L.A.; Benureau, F.C.Y. Sustainable Computational Science: The ReScience Initiative. *Peer J. Comput. Sci.* **2017**, *3*, 142. [CrossRef]
16. Crook, S.; Davison, A.P.; Plesser, H.E. Learning from the past: Approaches for reproducibility in computational neuroscience. In *20 Years in Computational Neuroscience*; Bower, J.M., Ed.; Springer: New York, NY, USA, 2013; pp. 73–102. [CrossRef]
17. Plesser, H.E. Reproducibility vs. replicability: A brief history of a confused terminology. *Front. Neuroinform.* **2018**, *11*, 76. [CrossRef]
18. Peels, R.; Bouter, L. THE possibility and desirability of replication in the humanities. *Palgrave Commun.* **2018**, *4*, 95. [CrossRef]
19. Leek, J.T.; Peng, R.D. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1645–1646. [CrossRef]
20. Penders, B.; Janssens, A.C.J.W. Finding Wealth in Waste: Irreplicability Re-Examined. *BioEssays* **2018**, *40*, 1800173. [CrossRef]
21. Devezer, B.; Nardin, L.G.; Baumgaertner, B.; Buzbas, E. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE* **2019**, *14*, e0216125. [CrossRef]
22. Peels, R. Replicability and replication in the humanities. *Res. Integr. Peer Rev.* **2019**, *4*, 2. [CrossRef]
23. Peels, R.; Bouter, L. Humanities need a replication drive too. *Nature* **2018**, *558*, 372. [CrossRef]

24. Peels, R.; Bouter, L. Replication Is Both Possible and Desirable in the Humanities, Just as It is in the Sciences. *LSE Impact Blog*. 2018. Available online: http://blogs.lse.ac.uk/impactofsocialsciences/2018/10/01/replication-is-both-possible-and-desirable-in-the-humanities-just-as-it-is-in-the-sciences/ (accessed on 21 March 2019).

25. Collins, H. *Changing Order: Replication and Induction in Scientific Practice*; University of Chicago Press: Chicago, IL, USA, 1985.

26. Pinch, T. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. *Sociology* **1986**, *20*, 653–654. [CrossRef]

27. Shapin, S.; Schaffer, S. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*; Princeton University Press: Princeton, NJ, USA, 1985.

28. Derksen, M.; Rietzschel, E.F. Surveillance is not the Answer, and Replication is not a Test: Comment on Kepes and McDaniel, "How Trustworthy Is the Scientific Literature in I–O Psychology?". *Ind. Organ. Psychol.* **2013**, *6*, 295–298. [CrossRef]

29. Leonelli, S. Re-Thinking Reproducibility as a Criterion for Research Quality (preprint) 2018. Available online: http://philsci-archive.pitt.edu/id/eprint/14352 (accessed on 21 March 2019).

30. Collins, H. *Forms of Life: The Method and Meaning of Sociology*; MIT Press: Cambridge, MA, USA, 2019.

31. Collins, H. Reproducibility of experiments: Experimenters' Regress, Statistical Uncertainty Principle, and the Replication Imperative. In *Reproducibility: Principles, Problems, Practices, and Prospects*; Atmanspacher, H., Maassen, S., Eds.; Wiley: New York, NY, USA, 2016; pp. 65–81.

32. Lamont, M.; White, P. *Workshop on Interdisciplinary Standards for Systematic Qualitative Research*; National Science Foundation: Washington, DC, USA, 2008.

33. Elman, C.; Kapiszewski, D. Benefits and Challenges of Making Qualitative Research More Transparent. Inside Higher Ed. 2017. Available online: https://www.insidehighered.com/blogs/rethinking-research/benefits-and-challenges-making-qualitative-research-more-transparent (accessed on 4 June 2019).

34. Gerring, J. What Standards Are (or Might be) Shared? In *Workshop on Interdisciplinary Standards for Systematic Qualitative Research*; Lamont, M., White, P., Eds.; National Science Foundation: Washington, DC, USA, 2008; pp. 107–124.

35. Comaroff, J. *Notes on Anthropological Method—Mainly in the Key of E, Workshop on Interdisciplinary Standards for Systematic Qualitative Research*; Lamont, M., White, P., Eds.; National Science Foundation: Washington, DC, USA, 2008; pp. 36–40.

36. Wythoff, G. On Method in the Humanities. *Configurations* **2018**, *26*, 289–295. [CrossRef]

37. Strübing, J. Problem, Lösung oder Symptom? Zur Forderung nach Replizierbarkeit von Forschungsergebnissen. *Forschung Lehre* **2018**, *25*, 102–105.

38. Strübing, J. Soziologie in kriegerischen Zeiten. *Soziologie* **2019**, *48*, 143–152.

39. Leung, L. Validity, reliability, and generalizability in qualitative research. *J. Fam. Med. Prim. Care* **2015**, *4*, 324. [CrossRef]

40. Chang, H. *Is Water H$_2$O?: Evidence, Realism and Pluralism*; Springer: Dordrecht, The Netherlands, 2012.

41. Corsín Jiménez, A. *A Data Governance Framework for Ethnography v. 1.0*; Spanish National Research Council: Madrid, Spain, 2018.

42. Nosek, B.; Alter, G.; Banks, G.C.; Borsboom, D.; Bowman, S.D.; Breckler, S.J.; Contestabile, M.; Dafoe, A.; Eich, E.; Freese, J.; et al. Promoting an open research culture. *Science* **2015**, *348*, 1422–1425. [CrossRef]

43. Haven, T.; Grootel, L. Preregistering Qualitative Research. *Account. Res.* **2019**, *26*, 229–244. [CrossRef]

44. DFG. Replizierbarkeit von Forschungsergebnissen. Eine Stellungnahme der Deutschen Forschungsgemeinschaft. 2017. Available online: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf (accessed on 30 May 2019).

45. Collins, F.; Tabak, L. Policy: NIH plans to enhance reproducibility. *Nature* **2014**, *505*, 612. [CrossRef]

46. Fineberg, H.V.; Allison, D. National Academies of Sciences, Engineering, and Medicine. In *Reproducibility and Replicability in Science*; The National Academies Press: Washington, DC, USA, 2019. [CrossRef]

47. Irwin, A. Re-making 'quality' within the social sciences: The debate over rigour and relevance in the modern business school. *Sociol. Rev.* **2018**, *67*, 194–209. [CrossRef]

48. Fricker, M. *Epistemic Injustice: Power and the Ethics of Knowing*; Oxford University Press: Oxford, UK, 2007.