On the use of replications in history

A white paper[1]

Pim Huijnen and Pieter Huistra (Utrecht University)

With the help of Auke Rijpma

With special thanks to Floris Boudens, Thomas van Gaalen, Hannah de Korte, Chiara Lacroix, Maiah Letsch and Matthijs Sweekhorst[2]

Version 1.1
Utrecht, 29 August 2022

How to cite this white paper: P. Huijnen and P. Huistra, "On the Use of Replications in History," white paper (Utrecht 2022).

About this white paper

In this white paper we address the scholarly community of historians. We argue that replication, or in more precise terminology: reproduction, is possible in historiography, and that it deserves a place in historical practice. Historical reproduction will:

- Complement and improve existing mechanisms of historical quality control such as referencing and peer review.
- Help to uncover the rules guiding historical work and will, in turn, improve the degree of methodological articulation and sophistication of historians.

The most important condition for historical reproduction to work is that we should adapt to the disciplinary specificities of history, instead of simply copying existing forms of replication now circulating in, mainly, the social and biomedical sciences.

This paper will proceed in 6 steps:

1. To show why attention to replication in historiography is timely.
2. To discuss the different forms of replication the literature distinguishes between.
3. To translate the existing forms of replication to historiography and to focus on reproduction as the preferred form of historical replication.

[1] This white paper was generously funded by Utrecht University's Fostering Open Science Fund.
[2] As explained below: the actual reproductions in step 5 of this white paper were indispensable for developing its argument. However, the authors remain fully responsible for the content of this paper.

4. To formulate the ambitions for reproduction, in terms of quality control and epistemic consolidation, and situate it in the existing landscape of historiographical quality control.
5. To present the results of our own project on replication in history, demonstrating that historical reproduction is practically possible and raises important but difficult questions for the discipline.
6. To formulate recommendations for the historical discipline.

1. Why is replication in history timely?

A replication crisis?

Replicability and/or reproducibility - more on definitions in the next paragraph - are the foundation on which the self-image of large parts of the sciences rests. To quote the opening sentences of three recent articles on the topic: ‘Replication [...] is considered the scientific golden standard’ (Jasny et al. 2011), ‘replication is a central tenet of science’ (Klein et al. 2014), and ‘reproducibility is a defining feature of science’ (Open Science Collaboration 2015). Replication and reproduction have a double aim. First, they are a form of quality control. Their goal is to map the grey area of questionable research practices and, in extreme cases, to detect fraud. Second, they aim to assess and enhance the epistemic robustness of research results.

Despite their sacrosanct role in scientific research, replicability and reproducibility have not prevented a sense of integrity crisis that has been haunting science since a decade or two. The principle alone did not prevent scientific misconduct, as a number of high-profile cases, such as Jon Sudbø in medicine, Jan Hendrik Schön in physics and Diederik Stapel in social psychology (David J. Hand 2020; Steen, Casadevall, and Fang 2013; Drenth 2015) have demonstrated.

These cases caused larger concerns over research integrity in general, as Stanford medical professor John Ioannidis voiced in his landmark article 'Why Most Published Research Findings Are False' (2005). He contended that replicability—in the sense of the transparency needed to redo a study-–in itself will not prevent questionable research practices like bias, cherry picking, and other types of questionable research practices. Results will (only) become more robust, he claimed, once different researchers ('dozens of them') study the same or similar questions. 'Unfortunately,' he continued, 'in some areas, the prevailing mentality until now has been to focus on isolated discoveries by single teams and interpret research experiments in isolation.' In other words: the principle of replicability (or reproducibility) should be followed up by the practice of replication (or reproduction).

The concerns that Ioannidis voiced, triggered numerous initiatives to assess the robustness of existing scientific practices. Especially in disciplines such as psychology and biomedicine, a series of systematic replications yielded some disturbing results. Many replications of papers that had been published in top-tier journals seemed to lead to different outcomes (Open Science Collaboration 2015). Equally disconcerting were the outcomes of a 2016 *Nature* survey, which

brought to the fore that an overwhelming majority of the responding scientists thought that the current state of reproducibility merited the use of the term ‘crisis’ (Baker 2016).

The aim of this white paper is not to add to the debate whether science really is 'in crisis'. Until now, the scientific methodological self-examination has left the humanities, in which replicability as a principle of scholarly soundness is highly uncommon, largely untouched. We take the so-called crisis as an opportunity to take a closer look at the research practices in the field of history. In that sense, we are now where many scientific fields found themselves almost ten years ago: to see what happens once the field of history starts to replicate existing studies. We have no idea what to expect. Even less, really, than those fields, since the question whether historical scholarship is reproducible is a much more open one—as is the question how to perform replications and/or reproductions in history. We will address these points in this white paper.

The double challenge of digitization

Is it problematic that we do not, yet, know what to expect from replications in history? We think so. What will happen once historians try and replicate each other's work is an urgent question to address. Already, digitization and datafication are impacting existing traditions in historical scholarship of quality control and trust.

Many historians have fundamental objections against replications. More on that below. However, practical difficulties have traditionally been just as important reasons for replications to be so uncommon in historical scholarship. Until recently, for example, it took a large effort (and expenses) to physically check references to archival sources. Digitization turns scrutiny of the provenance of certain claims into something as easy as clicking on a link in a reference. Making historians aware of this is the first challenge of digitization.

This is, on the one hand, true for the secondary literature that original research builds upon. The amount of ebooks and journals in our digital libraries surpass our physical ones by an arm's length. On the other hand, more and more of the primary sources historians base their scholarship on also can be found in online repositories. State archives are gradually digitizing their inventories. While it is true that only a fraction of all records in state archives can be found online, the process of digitization prioritizes much used archives that, in turn, receive an even higher chance to be referenced in historical literature. After all, digitization has an obvious pull factor for historical research. It is already noticeable how much scholarship flocks around massive digitized corpora of books (like Early English Books Online) or newspapers (like Chronicling America, the British Newspaper Archive, Gallica, Trove, Anno, or Delpher). Some of these collections are behind paywalls, but there are no technical reasons not to reference these repositories in publications directly.

To go one step further: the reasons not to share *any* source material historians have based their publications on in online repositories are–from a technical standpoint–rapidly becoming obsolete (obviously, there still are important issues of intellectual property rights and ethics to address,

but these go beyond the scope of this paper). Most historians digitize their primary sources themselves by taking pictures. A growing number of data management infrastructures can be used to share these with peers.

Making data available is, already, standing practice in areas of historical research that have a tradition of putting together datasets in the form of databases or spreadsheets. Often, these fields lean close to the social sciences, where data sharing is an intrinsic part of quality control. The same could be true in the foreseeable future for areas of history that are not, yet, used to even think in terms of 'datasets'. If this is so, it will only be a matter of time before historical scholarship will increasingly become subject to replication attempts, regardless whether we want them or believe in them. Again, it is worthwhile to start thinking about replications *before* this happens.

The second challenge is the increase of interdisciplinarity that digitization causes. Large-scale digitization of cultural heritage material has already prompted the interest of researchers from fields that have no intrinsic interest in historical perspectives. Now, computer scientists are studying the history of popular music in the USA, and the nature of conceptual change over time, while neuroscientists publish on the rise of social trust in Europe since the 16th Century (Mauch et al. 2015; Hamilton, Leskovec, and Jurafsky 2016; Safra et al. 2020)**.**

Consequently, significantly more historians than before will share source material, analytical techniques, and research interests with researchers from non-humanities fields. This will only proliferate once born-digital material like websites or social media reach the age to be considered historical sources. Quality control under these circumstances will involve sharing data and being explicit about methods as much as being transparent about the provenance of information from primary or secondary sources. What may be transparent for peers with a similar expertise might, after all, be questionable for readers from other fields.

In sum

The current focus in the sciences on replication and, more generally, on the ethics of research and transparency provides sufficient reasons to put this topic on the agenda in the field of history. The way the mass digitization of cultural heritage material paves for replication of historical scholarship only adds to its urgency.

However, replication in historical scholarship is not only an urgent issue to address, but also an important one in the sense that it can teach us many crucial lessons about the epistemology of historical research, which often remains rather elusive if not idiosyncratic—even for historians themselves. This will become clear, so we hope, from the remainder of this white paper.

There are obvious reasons why replication, thus far, has received little attention from the historical profession. The tenet of reproducibility and replicability, so central to the sciences, are commonly regarded a poor fit for the discipline's self-image as a field of study characterized by subjectivity and interpretation. And even if replications are deemed possible, the actual re-doing

of existing scholarship as a form of quality control lacks recognition and reward in an academic system that values novelty and originality. We contend in this white paper that replications, if properly adapted to the specifics of the field, are not only possible in historical scholarship. They can also be a valuable instrument to improve scholarly conduct and the field should, consequently, reward them as such. Before elaborating on this from paragraph 3 onward, however, we will address the question of what we talk about when we talk about replication.

2. What is replication?

We understand replication to be the re-performance of earlier research. Although scholars agree that there are different manners to do so, there is no consensus on nomenclature. Here, we follow Rik Peels' distinction (Peels 2019):

- Direct replication is the 'replication with a new data collection and with the same research protocol and the same research question as the original study'
- Conceptual replication refers to a replication with new data as well as a new research protocol, still with the same research question
- Reproduction is the exact repetition of earlier research: same research question, same research protocol, existing data sets

What distinguishes these three types, apart from the question whether or not they are based on the same or different data and/or methods, is their intended goal. The most important difference is between what here is called replication and reproduction. The goal of both conceptual and direct replication is to evaluate the results of the original study. The more different ways (in terms of data and/or methods) lead to corroboration of the original result (like 'a pencil in your mouth makes you feel happy', or 'vegetarians are less aggressive than meat eaters'), the more robust these results become. These results, we then tend to say, have become replicable (which, incidentally, these two examples are not). Reproduction adds another element: quality control. The goal of reproduction is not only to substantiate the results of the original study but also to evaluate its research design. Is it clear and convincing how the original research got from A (the research question) to Z (the conclusion)? Reproduction is then both a form of epistemic consolidation and a type of quality check in terms of transparency, adequacy, and plausibility.

	Scholar	Research question	Primary sources	Methods	Goal
Conceptual replication		✓			Epistemic consolidation
Direct replication		✓		✓	Epistemic consolidation

Reproduction		✓	✓	✓	Epistemic consolidation + quality control

In all three varieties two meanings of replication are conflated. One is replication as an activity - the re-performance of earlier research, the other is replication as an outcome - the fact that the re-performance of an earlier research leads to the same outcome as the first research. The normative expectation behind replication projects in the past few years is that replication in the first sense should lead to replication in the second sense - if we replicate research, we should see replication of the outcomes. If not, something is off. However, scholarship in the history of science has shown that this may be quite a lot to ask from existing research. Such an expectation supposes the possibility of complete transparency and replicability of existing research. As Steven Shapin and Simon Schaffer, and Harry Collins have famously shown for seventeenth-century air-pumps and twentieth-century lasers respectively: each aspect of scientific research involves a lot of tacit knowledge (Shapin and Schaffer 1985; Collins 1985). Even with elaborate sets of instructions and methodology sections, it turns out very difficult, to say the least, to rebuild a scientific instrument or re-perform an experiment. In these descriptive readings a replication is not a given implied in every sound research, but the result of a long process of negotiation. These insights from the history of science help to understand why 'failed' replications so often lead to discussions between replicators and original researchers whether the replication is a replication at all (e.g. Stroebe and Strack 2014) and they should make us cautious about the limits of replication and replicability.

3. How can we translate replicability to historiography?

On the limits of replication in historiography

Scholarly debates on the possibility and desirability of replications have thus far been discussed for the humanities in general, not for historical scholarship in particular. The focus in this debate has been on epistemic consolidation or, more specifically, on the question whether the outcomes of humanities' research should in principle be replicable and, thus, be allowed to be confirmed or rejected by new research. Rik Peels and Lex Bouter, most notably, have defended this position. They have pleaded for a 'replication drive' in the humanities in a letter to *Nature* (Peels and Bouter 2018a) and have subsequently elaborated their position in a number of articles (Peels and Bouter 2018b; Peels 2019). Their argument is clear and simple: the empirical disciplines in the humanities should be subjected to the same epistemic procedures as other empirical disciplines, 'namely that of replicating an original study to assess the likelihood that the original results are correct' (Peels and Bouter 2018b). Their expectation is that if research outcomes are correct, a conceptual or direct replication would lead to the same outcomes. Hence, Peels and Bouter argue that the humanities have to start replicating, because only the

practice of replication can teach the humanities whether they are subject to a crisis of unreplicable research results as well.

Sarah de Rijcke and Bart Penders have defended a diametrically opposed position towards replication in their reaction to Peels and Bouter's letter in *Nature* (De Rijcke and Penders 2018). De Rijcke and Penders argue that the humanities should 'resist' the call for replication, because it departs from the wrong epistemic assumptions about 'correctness' of research outcomes in the humanities. Instead of a single correct outcome to a given research question, the humanities allow for 'the coexistence of multiple valid answers' (De Rijcke and Penders 2018). This is caused by the interpretative nature of the humanities, which 'pursue meaning beyond truth', and thus allow considerable more leeway than a binary notion of correct versus incorrect.

Both sides of this debate share an understanding of replication that has as its main goal the assessment of the correctness of outcomes. Replication of this kind would indeed be a novelty for historiography. It use however - and here we side with De Rijcke and Penders - would be very limited, namely to the very trivial truths that have been confirmed time and again, such as 'The First World War started in 1914'. In historiography's self-understanding, things rather work the other way around: what is interesting is not what historians agree, but what they disagree upon. Take a well-known example of a question that has been studied widely for over a century now, 'What caused the First World War?'. Historians have approached, and continue to approach, this question from a number of angles. They have added visual or material sources to a more traditional focus on documents; they have substituted the political for a more socio-economic or cultural outlook; they have widened the geographical scope from Europe to a more global perspective. Historians have, in replication terms, applied different methods and different data to the same research question. In this sense, the historiographical debate meets some of the requirements of direct and conceptual replication, so one could argue that historians already perform replications. However, they are not replications in the sense propagated by Peels and Bouter, because they do not seek to check whether the results of earlier historians' research are correct or not. Rather, they add perspectives, enrich the understanding of the causes of the First World War and thus achieve what progress amounts to in historiography. To historiography applies what De Rijcke and Penders mean when they stress 'the coexistence of multiple valid answers and the value of their interaction' (De Rijcke and Penders 2018).

Of course it is questionable if it is justified to equate historiographical debate with replication, but this only shows the difficulties involved in translating the idea and practice of replication from the biomedical and social sciences to, in this case, historiography. Indicative of this as well, is the use of 'data' by Peels and Bouter. The notion of 'data' is alien to many, but not all, humanists. Historians, for example, conceive of their raw materials as 'primary sources'. The difference between 'data' and 'sources' and their handling, makes it unclear what, for example, is meant by using 'existing data'. Would historians have to go to the same archive and read and interpret the same sources as the colleague they aim to replicate? Or would they need to use the notes already made and treat them as data? What gets lost in translation between data and source is

more than a practicality; it points to the fundamental need of building scholarly practice on disciplinary specificities.

Minimal reproduction as historiographical replication

Precisely this need for disciplinary specificity has been argued for by a number of commentators, and has been termed ‘new localism’ (Guttinger 2020). Such localism has also been forwarded by Penders and De Rijcke, in their follow-up article on replication in the humanities (Penders, Britt Holbrook and De Rijcke 2019). Replication may be laudable in many fields, they argue, but not in the humanities. The implementation of replication and replicability in the humanities would mean imposing ‘an imagined unified, monistic science’ and overlook epistemic differences. Sabina Leonelli (2018) has also argued against such a one-size-fits-all approach. Her approach is different to the extent that she has tried to define a number of different forms of reproducibility, ranging from computational reproducibility in software research to principally irreproducible forms of research, such as participant observation in anthropology.

To some extent, our argument follows Leonelli’s line here: we argue for a form of historiographical replication that takes into account disciplinary specificities. To find such a form we have to move beyond the focus on epistemic consolidation that both positions in the debate on replication in the humanities seem to hold as the central and sole aim of replication. Reproduction, which focuses on quality control as well as epistemic consolidation, is the obvious candidate for a fitting form of historiographical replication. After all, is not the footnote already an invitation to perform such a reproduction? It is in their footnotes that historians show where their claims and interpretations come from and make themselves accountable. Footnotes offer the opportunity to follow historians' footsteps.

As we noted above, reproduction is about following the footsteps of earlier historians, but it also entails the expectation of epistemic consolidation. Namely the expectation that if one historian follows the footsteps of another, they would end up with the same answer to the same research question. However, is that not an unrealistic expectation? After all, are the interpretations historians make not necessarily underdetermined by the evidence that is offered for them? And even more importantly: does a plea for historiographical reproduction not overlook the fact that there is still one important difference between original research and reproduction, namely the scholar? If historiography is a deliberately subjective discipline, in which the person and the background of the scholar, rather than a hindrance, are a necessary precondition for acquiring knowledge, then no two scholars, except perhaps identical twins, can be expected to produce the same outcomes.

These objectives are valid, and explain why we cannot opt for a *maximal* reproduction, by which we understand following an earlier historian from A (research question) with the expectation to end up at Z (the answer to that question). Such an understanding of reproductions is prevalent in other disciplines and seems to have informed both sides of the debate on replication in the humanities. This would be too much to ask from historiography. That is why we propose a *minimal* reproduction, in which a reproduction would start at Z, and then work iets wat back to A.

The requirements for such a minimal reproduction are much less strong: the question at stake is not whether the same research would lead to the same outcome, but whether the given outcome is sustained by the sources and methods used in the original study.

Such minimal reproductions have thus far not been performed systematically in historiography. The work by Hallie Lieberman and Eric Schatzberg on Rachel Maines' *The Technology of Orgasm* (1999) can count as a rare exception, one that immediately proves the validity of such an approach (Lieberman and Schatzberg 2018). Lieberman and Schatzberg scrutinized Maines' acclaimed work and their well-documented conclusion was that its main claims did not hold. Lieberman and Schatzberg followed Maines' footnotes and consulted the sources she referenced to back up her claims. In these sources, Lieberman and Schatzberg simply could not find evidence for the fact that doctors in the Victorian age had routinely used vibrators for clitoral stimulation, to induce an orgasm as a form of therapy for patients who had been diagnosed as hysterical. Lieberman and Schatzberg, thus, showed that, although multiple valid answers in historiography may coexist, their number is not infinite. There *are* unvalid answers. They presented their own work as the exposure of a 'failure of academic quality control'. In terms of this white paper, they have also proven that quality control in the form of reproduction is possible.

4. How does reproduction fit in with existing forms of quality control and what may we expect from it epistemically?

Reproduction in relation to trust-establishing practices

Let us not be misunderstood: reproduction is not the only form of quality control, nor will we suggest in this white paper that it should somehow replace existing forms of quality control. Rather, weak reproduction can be understood as an addition to an already existing regime of quality control. Mechanisms such as the above-mentioned footnotes, the historiographical debate and peer review have been in place for a long time in historiography. Footnotes are meant to make claims traceable, the historiographical debate ensures a critical discussion between members of the discipline, while peer review fulfills a more or less similar function as a gatekeeper for articles, books and grant applications. Together they form a landscape of what has been called 'trust-establishing practices' (Guttinger 2020).

Trust is then something that is caused by these practices, but it is something that at the same time is a prerequisite for these practices - as the difference between reproduction and these existing practices makes clear. Reproduction starts from a different ambition than historiographical debate: it does not seek to add a new viewpoint to the debate, but to evaluate an already existing contribution. Reproduction goes beyond peer review, because a peer reviewer typically reviews the relevance of an article's research question and the value and

urgence of its contribution to the field rather than the degree to which an author is transparent. Footnotes, finally, are usually taken on trust, and not systematically and exhaustively checked.

Reproduction does not take this trust - that footnotes are accurate or that the existing contributions to a historiographical debate are solid - for granted. In this sense, reproduction is not another trust-establishing practice, but a form of control of these practices. As such, we could see reproductions as a form of systematic distrust. In turn, if the results are reassuring (which in the case of Maines they were not) this systematic distrust can lead to an increase in trust in historical scholarship. Do note here, that reproduction-as-distrust can only 'take place on the margins of trusting systems' (Shapin 1994). The work by Lieberman and Schatzberg is a case in point: in their account of the work of Maines, they make use of the exact same trust-establishing practices as Maines herself, such as footnotes. Our choice to favor Lieberman and Schatzberg's account over Maines' is then, in final instance, not based on our own reproduction of Maines' work (or of our reproduction of Lieberman and Schatzberg's reproduction of Maines' book), but on our assessment of the greatest trustworthiness. This point immediately makes clear that a plea for reproduction is never a plea *against* trust itself, because this would lead to infinite regression of reproducing reproductions.

The epistemic return of minimal reproductions

Minimal reproduction is about the systematic retracing of a historian's argumentative steps, in a backward direction. It is important to stress the *systematic* character of such retracing: it is not about making a carbon copy of the practice of the earlier research. Such a copy of the practice - the *context of discovery*, philosophers of science would say - would include all the detours, the chance finds, and eureka moments that are so characteristic of any research process. Such aspects are probably among the most charming aspects of historical research, but they do not deserve a role in the foundations of a historical claim. Johan Huizinga's famous thesis on the fourteenth- and fifteenth-century Burgundian culture as a demise (an autumn) rather than as a new dawn, came to him during one of his walks outside of the city of Groningen. Such an epiphany may prompt a new historiographical thesis, but it can never serve as its justification. And it is in this *context of justification*, the systematization of the messy practice of research, that we have to situate the work of reproduction.

That we are interested in the justifications of the historian's claims, makes clear that in minimal reproduction the stakes are always also epistemic, or cognitive, and not only about quality control. Once again: the aim of historiographical reproduction should not be to see if repeating earlier research leads to the same outcome. That we do not believe in one correct and repeatable outcome of historical research, does not mean that anything goes in terms of reaching such outcomes - there are rules for doing historical research.

Recently, the philosopher of history Jouni-Matti Kuukkanen has forwarded the idea to see historiography as a form of 'claiming correctly'. Historiography, in Kuukkanen's reading, should not be treated by the standards of immediate empiricism. There is no way in which we can assess the historian's claims by comparing them with past reality, since our access to that past

reality is always mediated through evidence - what historians commonly refer to as ‘sources’. So, instead of seeing the historian’s text as a representation, a mirror-image, of the past, we should treat it as a claim, or set of claims, that are based on inferential reasoning from the available evidence. Such an understanding of historical claims leaves no room for assessing their *truth* in terms of the correspondence between text and reality. What it does allow, is passing an epistemic judgment about the *correctness* of these claims. A claim should be treated as correct if it is the result of sound inferential reasoning. (Kuukkanen 2020).

Kuukkanen’s reading of historiography offers an important and usable tool for performing historical reproductions. If the historian’s work consists of making claims that are based on inferential reasonings, then in a reproduction we may check whether the historian’s inferences are sound. In other words: whether they follow the historiographical rules. But this, obviously, begs the question of what these historiographical rules might be. What, exactly, distinguishes a valid interpretation from a invalid one? Reproductions offer an outcome to this question: they unearth the rules of inference that are actually prevalent in historiography. As we will further elaborate below, the epistemic yield of reproductions thus was twofold: in reproducing the justifications historians make for their claims, we were able to identify the rules historians use for these justifications. Kuukkanen, himself a philosopher, for this reason has called for a thorough study of historiography, ‘to study the nitty-gritty and follow inferential chains and networks’, to see how these rules actually function (Kuukkanen 2020). Our experiments in replication are an answer to this call.

5. To what results does historical reproduction in practice lead?

Once more, with feeling

In early 2021, when we started our experiments in replication with our research project ‘Once more, with feeling. Replication to improve open knowledge production in the humanities’ (funded by Utrecht University’s Fostering Open Science Fund), we did not have these clear conceptual distinctions. We did not make a clear distinction between the two possible aims of replication, assessing epistemic robustness and quality control. Let alone that we had already separated minimal from maximal reproduction. We simply started out from the assumption that some form of reproduction was implied in the disciplinary ethic through footnotes; that reproduction had not been tried and tested in a systematic way in history; and that the best way to find out its merits and disadvantages would be to do just that. So, that is what we did in the project. It is important to note that our purposes in this project have not been to analyze existing attempts at replications, but to try them out ourselves. Our understanding of replication, therefore, is normative and prescriptive.

When it comes to the actual performance of these experiments in replication, it is important who we mean by ‘we’. We, as in the author of this white paper, were the project leaders. As such we

are responsible for the design of the project, its supervision and realization. The actual work of the replications, however, was not done by us, but by six graduate students: Floris Boudens, Thomas van Gaalen, Hannah de Korte, Chiara Lacroix, Maiah Letsch and Matthijs Sweekhorst (at the time of this project they were all Research Master students in history at Utrecht University). As graduate students they were trained in historical methods, but not yet completely encapsulated in the peculiarities of historical practice, which allowed them to approach the subject matter with an open mind. The six students formed three pairs of two students each, which worked under our close guidance. We had regular project meetings (every 2-4 weeks) and smaller team meetings in between.

Each of the groups set out to reproduce one article, one from the respective subdisciplines of the project supervisors: digital history, cultural history and economic history. These three subdisciplines also offered the widest possible variety within historiography, among other things in terms of their natural affinity with reproduction. Our suspicion was that both the method-centered digital history and the quantitative and objectivist approach of economic history would be an easier fit for the idea of reproduction, contrasting with the subjectivist and classic hermeneutical nature of cultural history.

In the project we worked in three consecutive steps. First, we tried to find an approach to historical reproduction, followed by the search for a suitable article to reproduce. Second, we developed a method to actually perform the reproduction. Third, we performed a reproduction. Below, we will elaborate on these steps and present our most important findings.

Selection

Selecting an object for reproduction was not as easy as we initially expected. Due to time constraints we immediately left aside book-length historical work and opted for article-length publications instead. We formulated criteria for suitable ones. Obviously, the articles had to be from the correct subdiscipline. Moreover, the publications needed to have some renown within the subdiscipline, or in any other way had to be exemplary for the current state of the subdiscipline, so as to guard ourselves against criticism of having chosen atypical or exceptional articles. Finally, the articles needed to be (seemingly) reproducible. We demanded both a substantive empirical analysis and an elaborate explanation of the manner in which the research was performed (through footnotes, methodology sections, et cetera).

The first insight that our project produced was that, following these criteria, not many articles are at all suitable for reproduction. An important reason for this is that the empirical nature of scholarly output can be highly dissimilar. Many publications focus primarily on theoretical or conceptual issues, which makes them less suitable candidates for reproduction. Of course, there is no clear-cut distinction between the empirical and the theoretical, since they often inform each other, nor does the lesser reproducibility of theoretical articles turn them into unsound scholarship. However, it did make our search for suitable articles to reproduce less trivial than foreseen.

More problematic were those articles that did not allow for their reproduction because of insufficient referencing or too little information on research methods. The latter was mainly a problem in digital and economic history, because method has a more central position in these fields. Details on the exact data and techniques are simply indispensable to even start reproducing research in these fields. By contrast, cultural history has a tradition of not elaborating on methodology at all. We will return to this point below. However, it is important to underline that, given these practical limitations, it was not difficult at all to find unreproducible publications. However, proving this was not the point of our project. We were eager to retrace the steps of a study that *did* seem reproducible and, therefore, settled for publications that looked as transparent in terms of method as possible.

The articles we, eventually, tried to reproduce were: Vilja Hulden, 'News Diets: Main Courses and Side Dishes', published in 2020 in *Current Research in Digital History*; a cultural historical article published in a leading historical journal;[3]; Guido Alfani, 'Plague in Seventeenth-Century Europe and the Decline of Italy: an Epidemological Hypothesis', published in 2013 in *European Review of Economic History*. Hulden's article offered empirical material, recent digital historical work and an elaborate methodology. The article from cultural history came from an established domain of cultural history, the history of sexuality, and it was the most read article from the journal in question . Alfani's paper was well-cited and offered enough data to be reproducible (while, not unimportantly, its quite technical methods fell within our skillset).

Developing a method

After the selection of the articles, we developed a method for historical reproduction, since no such method was yet available. It soon became clear that simply checking all the footnotes would not be feasible within the allotted time, nor would every footnote be of equal interest. Footnotes have many functions, and not all of them are helpful in the process of reproducing the historian's work. There are footnotes that serve as signs of erudition, by citing standard works from the field for example, and footnotes that function as bibliographical signposts to help other historians on their bibliographical quests. Only following footnotes would, thus, be too wide and too narrow at the same time. After all, not all that is important in an article is referenced in footnotes.

For these reasons, we started our reproductions from the interpretations offered in the article. This required an act of translation to make the article into a reproducible unit. This translation entailed the identification of, first, the main claims of the article and, second, the dozen or so supporting claims. These claims were put in a hierarchy. Following from this, it was possible to

[3] We do not discose the title and author of the article at this point, because it has recently come to our knowledge that the scholar who wrote the article does not have a permanent position at a university. We were, wrongly, convinced that this was the case. Although our aim was not to check the merits of the work of one scholar, but rather to find out the state of the (sub)dscipline, we do understand that our approach could be misread as such, with possible negative consquences for the historian in question. Since temporary staff are in a precarious position in academia, we think it is justified for now not to disclose the identity of the article and the author. This does not mean the findings from the replication are no longer valid. Therefore, readers who are interested in reading the underlying report are invited to address the authors of the white paper through email.

identify the relevant footnotes, and thus the primary and secondary material on which these claims rested. This procedure worked well and gave us a hands-on method for reproduction, because it made clear what it was that needed to be reproduced. Unknowingly at the time, we chose an approach that aligns very well with an understanding of historiography as ‘claiming correctly’.

There are at least two caveats to this claim-based approach to reproduction. First, this act of translation obviously requires an intervention from the replicator. Whether the identified claims are the ones authors themselves would have chosen, or if another replicator would judge differently, is up for debate. Second, this translation favors a certain type of history writing, namely the type of history writing that makes clear claims, that can be found in abstracts and conclusions. Every form of history writing, to a greater or lesser extent, makes claims. However, our understanding of reproduction obviously targets the journal article in its current form much better than, for example, a synthesis or a chronicle.

A procedure for reproductions in history

Translated into a more schematic form, our procedure for reproductions looks like this:

1. Keep a reproduction protocol
2. Identify
   a. the central claim of the study (research question + conclusion)
   b. the research protocol of the study (method, practical steps)
   c. the source material the study is based on
3. Make explicit the strain of the argument in the study from claim to question, identify all the steps that are explicated
4. Assess the extent to which each step can be traced back to secondary literature, primary sources, or output from quantitative/computational techniques. The questions here are always:
   a. is the claim plausible in the light of the material (literature, sources, output) used?
   b. Is it clear (transparent) how the author has deducted the subsequent interpretative step from the former?
5. Write reproduction report

Findings

Based on this procedure for reproduction, what did we find? Most importantly, we found that **reproductions in history are possible**, but also that they are difficult to perform.

Reproductions are possible in the sense that it turned out to be feasible to retrace the steps of the historians who wrote the three articles, and to assess whether their interpretations were sustained by the underlying research. At the same time, it turned out that reproductions were quite a difficult and time-consuming activity, because the existing articles did not offer a fully transparent step-by-step guide for reproductions. It was during the reproductions that we found

out that we could make a distinction between minimal reproduction, working backwards from claims to research, and maximal reproduction, working forwards from an existing research protocol. **Minimal reproduction proved the possible and better-suited option for historiography.**

Right from the outset, it was clear that the cultural-historical article would not provide the possibility for a maximal reproduction. In line with prevailing practice in cultural history, there was only a very short clause on method (identified as 'close discursive analysis' but not defined any further) and the body of sources was not clearly delineated: this could hardly serve as a recipe for reproduction. Hulden's article, however, seemed to promise exactly that: a clear explanation of the programs and code used to do research, files added with the goal of transparency in mind and an accessible dataset. Hulden's article thus seemed perfectly (maximally) reproducible – until it wasn't. Crucial files for the type of topic modeling used in the article were missing, and certain steps of interpretation of the results remained elusive. A maximal reproduction turned out to be impossible here as well.

The failure to maximally reproduce Hulden's article was an important finding in itself. It proved to us that the availability of ample footnotes and access to primary sources (or data) and the techniques used (like code) in itself are not sufficient markers for reproducibility. Even if a study may seem entirely transparent, **you need to perform a reproduction to find out to what extent an article actually is reproducible**. Reproduction, thus, provides us with a means to assess historical reproducibility. And it pointed us in the direction of minimal reproduction as a suitable form of reproduction for the discipline of history. The limited extent of Hulden's reproducibility, we assess, was not contingent on a missing file or on the outsourcing of reprint detection to other scholars. Rather, it laid bare the specificity of historical research, with its many steps of interpretations, its choices from different methods, its highly subjective selection of source material – all this makes minimal replication preferable over maximal replication.

Reproduction as a means of assessing reproducibility may sound as a less exciting and more circular conclusion than it actually is. However, assessing to what extent a certain historical text allows for a reproduction amounts to showing how historians justify the outcomes of their work. In this sense, a reproduction opens up the black box of historiography. It lays bare what steps of the historical work remain opaque, such as the steps Hulden took in the interpretation of her topic modeling, or the cultural historian's' reliance on secondary source material. Also, if such steps in the production of historical claims are not included in a historical text, then the replicator has to try to infer what they must have been. **What reproductions provide us with, therefore, is a deeper understanding in how historians justify their claims, as well as what they omit.**

Regarding the latter, the omissions, we conclude on the basis of an admittedly very small sample of three articles, that quite a few of the interpretative steps historians make remain unclear or unsaid. **Historical work turns out to be intrinsically intransparent**: methodologies remain unarticulated, interpretations are not elaborated upon, the origins of quantitative data

remain unclear, et cetera. Of course, this is not necessarily the case for all historical work, but it is our belief that we identified disciplinary conventions here, rather than three outliers.

Reflection on our findings

That the historical work in our experimental replications proved to be not fully transparent merits some further reflection. First of all, one could argue that we have relied too much on the historical text only, and that if we want to know how historical claims come into existence, we should have equally focused on actual historical *practice*. It, probably, is the case that historians give more thought to, for example, the way they interpret their sources than they have words available in their papers. However, it is our belief that in published articles we should not only find the outcomes of research, but also a justification of how these outcomes have been reached.

Second, transparency is a notion that should not be understood in a binary fashion. It is not the case that historical (or other scientific) texts are either fully transparent or entirely intransparent. Rather, there are differing degrees of transparency, and, as already mentioned above when discussing Collins' work on replication in the sciences, there will always be an element of tacit knowledge when it comes to interpretation. However, this does not exonerate historians from aiming to be as clear as possible about what they do when they make historical claims. And, according to us, transparency in historical work should be improved.

Third, we are not trying to blame any individual authors here. **Transparency is a shared responsibility.** It might not always be the author's choice *not* to give an elaborate explanation of how they interpret the outcomes of their topic modeling, or of what they understand close discursive analysis to be. Rather, the status of method in scholarly output reflects disciplinary habits. These articles were all published by academically employed scholars, went through peer review and the scrutiny of an editorial board. Consequently, they represent not the idiosyncrasies of a few authors, but the norm within our discipline. That norm extends to our own work, that, if replicated, would probably be subject to similar criticisms. Nevertheless, it is our belief that we as the historical discipline can and should do better.

6. How to proceed with reproduction in historiography?

Experiments in historiographical replication should not end here. We have four recommendations (accompanied by three warnings).

First, **we need more reproductions based on a stricter protocol than the one we employed in our project**. Our protocol was made up as we went, so this would merit further and more systematic testing. The five steps mentioned in the previous section serve as an invitation and, of course, leave room for improvement. Also, further reproductions will allow us to assess the representativeness of our findings. At the same time, we do not hold a plea for endless

reproductions. **Not every historical article should be reproduced**. As we explained in section 4: it is not our aim to replace trust with distrust.

Second, to increase the number of reproductions, **we need to provide more incentives for reproductions**. A practical reason that reproductions in history are non-existent is the fact that they lack status as accepted scholarly work. Reproductions do not count as research, nor as publications. There are relatively easy ways to change this. We envisage funding instruments for reproductions as well as the possibility of publishing the results of reproductions, regardless of their outcomes. However, **this should not lead to a separate subfield of reproduction studies**. We think it is much more promising if reproductions become part of existing practices and existing publication structures, rather than becoming a niche of its own. Reproductions, as we have experienced in this project, seem especially suitable for graduate training. It is for this reason that in 2023 we will offer a course in reproductions in cultural history at the Netherlands Research School for Cultural History (Huizinga Institute).

Third, **we need to increase the reproducibility of historical work**. To do so, we need a more elaborate and better methodological articulation in the sense of a clearer explanation of what we actually do when we make claims. It is important to underline that this should not be misunderstood as a plea for adopting the 'scientific method' as practiced in many of the social sciences. **We do not want to turn history into a discipline that solely focuses on quantitative data and sees hypothesis testing in terms of statistical inference as the quintessence of scientific and scholarly activity**. Instead, we argue that history should hold on to its subjectivist self-understanding and should not lose its hermeneutical methods or narrative forms. What it *should* do is explain more and better what these things entail, exactly. In practice, such explanations need not be the obligatory methods sections seen in so many disciplines, but could easily form a second or third layer behind the narrative one that digital publishing allows for.

Our fourth and final recommendation is actually more of an expectation. We think that **the adoption of reproduction in historiography will increase the epistemic robustness of historical work**. What we do not mean is a simple true/false notion of historical claims, because such a notion does not fit with historical interpretations. Instead, we think of an increase in epistemic robustness through a clearer idea, articulation and execution of the rules of making historical inferences, that support these claims.

Bibliography

Baker, Monya. 2016. “1,500 Scientists Lift the Lid on Reproducibility.” *Nature News* 533 (7604): 452. https://doi.org/10.1038/533452a.

Collins, Harry. 1985. Changing Order. Replication and induction in scientific practice. Chicago: Chicago University Press.

De Rijcke, Sarah, and Bart Penders. 2018. “Resist Calls for Replicability in the Humanities”. *Nature* 560 (7716): 29–29. https://doi.org/10.1038/d41586-018-05845-z.

Drenth, Pieter J.D. 2015. “What Lessons Can We Learn from the Stapel Case?” In *Integrity in the Global Research Arena*, edited by Nicholas Steneck, Melissa Anderson, Sabine Kleinert, and Tony Mayer, 147–56. Singapore: World Scientific Pub.

Guttinger, Stephan. 2020. “The Limits of Replicability”. *European Journal for Philosophy of Science* 10 (2): 10. https://doi.org/10.1007/s13194-019-0269-1.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. “Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change”. *Proc Conf Empir Methods Nat Lang Process*: 2116-2121. https://doi.org/10.18653/v1/d16-1229.

Hand, David J. 2020. Dark Data: Why What You Don’t Know Matters. Princeton University Press. https://doi.org/10.2307/j.ctvmd85db.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Jasny, Barbara R., Gilbert Chin, Lisa Chong, and Sacha Vignieri. 2011. “Again, and Again, and Again ….” *Science* 334 (6060): 1225–1225. https://doi.org/10.1126/science.334.6060.1225.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. “Investigating Variation in Replicability.” *Social Psychology* 45 (3): 142–52. https://doi.org/10.1027/1864-9335/a000178.

Kuukkanen, Jouni-Matti. 2020. “Historiographical Knowledge as Claiming Correctly.” In idem (ed.). Philosophy of History: Twenty-First Century Perspectives. London et.al.: Bloomsbury Academic: 44-65.

Leonelli, Sabina. 2018. “Re-Thinking Reproducibility as a Criterion for Research Quality.” Preprint. January 28, 2018. http://philsci-archive.pitt.edu/14352/.

Lieberman, Hallie and Schatzberg, Eric. 2018. "A failure of academic quality control: The technology of orgasm". *Journal of positive sexuality* 4, 2: 24-47.

Maines, Rachel. 1998. The Technology of Orgasm “Hysteria”, the Vibrator, and Women’s Sexual Satisfaction. Baltimore: The Johns Hopkins University Press.

Mauch, M. et al. 2015. “The Evolution of Popular Music”. *Royal Society Open Science* 2: 150081. https://doi.org/10.1098/rsos.150081.

Open Science Collaboration. 2012. “An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science.” *Perspectives on Psychological Science* 7 (6): 657–60. https://doi.org/10.1177/1745691612462588.

Peels, Rik, and Bouter, Lex. 2018a. “Humanities Need a Replication Drive Too”. *Nature* 558 (7710): 372–372. https://doi.org/10.1038/d41586-018-05454-w.

———. 2018b. “The Possibility and Desirability of Replication in the Humanities”. *Palgrave Communications* 4 (1): 1–4. https://doi.org/10.1057/s41599-018-0149-x.

Peels, Rik. 2019. “Replicability and Replication in the Humanities.” *Research Integrity and Peer Review* 4 (1): 2. https://doi.org/10.1186/s41073-018-0060-4.

Penders, Bart, Holbrook, J. Britt, and De Rijcke, Sarah. 2019. “Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing”. *Publications* 7 (3): 52. https://doi.org/10.3390/publications7030052.

Safra, Lou, Coralie Chevallier, Julie Grèzes, and Nicolas Baumard. 2020. “Tracking Historical Changes in Trustworthiness Using Machine Learning Analyses of Facial Cues in Paintings.” *Nature Communications* 11 (1): 4728. https://doi.org/10.1038/s41467-020-18566-7.

Shapin, Steven. 1994. A Social History of Truth. Civility and Science in Seventeenth-Century England. Chicago: Chicago University Press.

Shapin, Steven and Schaffer, Simon. 1985. Leviathan and the air-pump: Hobbes, Boyle, and the experimental life. Princeton: Princeton University Press.

Steen, R. Grant, Arturo Casadevall, and Ferric C. Fang. 2013. “Why Has the Number of Scientific Retractions Increased?” *PLOS ONE* 8 (7): e68397. https://doi.org/10.1371/journal.pone.0068397.

Stroebe, Wolfgang, and Fritz Strack. 2014. “The Alleged Crisis and Illusion of Exact Replications”. *Perspectives on Pyschological Science* 9 (1): 59-71. https://doi.org/10.1177/1745691613514450.