Zygon®: Journal of RELIGION & SCIENCE

Brown, Jeremy. 2024. "Crisis, Confidence, and the Limits of Replication." *Zygon: Journal of Religion and Science* 59 (2): 575–90. DOI: https://doi.org/10.16995/zygon.11502

Open Library of Humanities

Crisis, Confidence, and the Limits of Replication

Jeremy Brown, Director, Office of Emergency Care Research, National Institutes of Health, Rockville, MD, USA, jeremybrownmd@gmail.com

There have been calls for a program of replication in the humanities. Although usually thought of as confined to the hard sciences, replication may, under the correct conditions, be a useful tool for historians who propose an explanation of why a set of events occurred. But the program of replication in the humanities is challenged when we consider degrees of freedom, i.e., the number of independent parameters that function within a system. Evidence from the sciences has revealed that experimental variables once thought of as unimportant might in fact be critical. Change just one of them and the experimental result changes in ways that were at first unimaginable. How then, are we to know if the degrees of freedom offered as part of a historical explanation are indeed satisfactory? There are constraints to what may be replicated, but this is the case for the sciences no less than for the humanities.

A Divided Court

In 2008, the Supreme Court of the United States ruled 5–4 that the Constitution protected the right of an individual to possess a firearm for protection within their home. This right, wrote Justice Antonin Scalia for the majority, was based on a reading of the Constitution's Second Amendment, in which an individual's right to possess a firearm was unconnected with service in a militia.[1] In a strongly written dissent, Justice John Paul Stevens opined that, in fact, "the Amendment is most naturally read to secure to the people a right to use and possess arms in conjunction with service in a well-regulated militia." The Supreme Court justices could not agree on the meaning of some words written only 250 years ago in a language in which they are both fluent, and with a legal training common to all. How the Constitution is read, and what its framers may have had in mind (and whether or not that is important), potentially has life-threatening consequences.[2] Historiographical disputes, less so. But, like justices of the Supreme Court, historians often disagree on the meaning of a document or a series of events. The historian Arthur Schlesinger Sr. (1888–1965), for example, wrote that the American Revolution was prompted by economics, not politics: "Less intent on politics than business, the merchants as a class did not ordinarily concern themselves with political questions. But when their interests were jeopardized they entered politics with a vim, and might be expected to carry things their own way" (Schlesinger 1957). In contrast, Daniel Boorstin (1914–2004) claimed that the American Revolution had only political goals: "The political objective of the Revolution, independence from British rule, was achieved after one relatively short effort. 1776 had no sequel and needed none: the issue was separation, and separation was accomplished" (Boorstin 1989).

In an effort to adjudicate between these and other hypotheses about history, there have been calls for a program of replication in which the evidence supporting differing models can be reexamined in an effort to establish their epistemic status.[3] This article addresses this program by comparing—and contrasting— it to recent replication efforts in the basic and clinical sciences. There is much to be gained from understanding the reasons for and approaches to replication in the sciences and delineating what can and cannot be reasonably expected from such efforts.

Before proceeding, I should note that the observations I make about history also apply in large part to many of the other disciplines that make up the humanities. Here, I follow the lead of Rik Peels and others who include archeology, history (and its various subdisciplines such as the history of science and the history of religion), linguistics, philosophy, theology, and religious studies, along with several other areas, as belonging to the humanities. Other disciplines, such as basic and clinical medical research, physics (though not necessarily theoretical physics), chemistry, and the biological sciences are included in the broad category of the sciences (Peels 2019, 11n).

As I discuss in some detail, the call for replication in the sciences was the result of a crisis. Many experiments, some published to great fanfare and media acclaim, could not be replicated, and the scientific enterprise urgently needed a correction. In contrast, the call for replication in the humanities seems to have originated from a position of confidence. While not always explicitly stated, calls for a program of replication in the humanities stem from the belief that the stakes for a theory in the humanities matter in a way similar to the stakes for a laboratory experiment or clinical trial. A useful example was suggested by Peels. In 1993, Samuel Huntington published *The Clash of Civilizations and the Remaking of World Order*, suggesting that future wars will be increasingly fought between ideologies rather than between countries. This premise, which was widely cited and discussed, had "important repercussions in cultural anthropology, history, peace and conflict studies, political theory, and theology and religious studies" (Peels and Bouter 2018), and, I might add, has critical implications for the allocation and prioritization of enormous defense budgets. The call to see if Huntington's hypothesis could be replicated did not originate from a position of crisis but rather from one of confidence. The hypothesis has real-world implications, and its replication was therefore not just an interesting intellectual exercise. It was a practical matter of considerable urgency.

I begin with a description of three types of history and explain what kind of replication may be expected for each. I then briefly describe replicating my own work in the history of science and religion, and then outline the recent replication movement in the sciences. I show that although there may only be limited opportunities for replication in the humanities, this is also a feature of replication in the sciences. The penultimate section introduces the statistical concept of degrees of freedom and how it presents a challenge to replication in both the humanities and the sciences. In my concluding remarks, I clarify the limits of scientism and the need for the humanities to demarcate what should and should not be replicated.

Some Remarks on the Nature of History

Within the humanities in general, and history in particular, what is the relationship between a falsifiable scientific hypothesis and the ideas of rigor and reproducibility? To answer this, it must be decided what reproducibility in history is attempting to do. To answer that, it must be decided what is meant by history. This is, of course, an enormous question, but broadly, there are three goals a historian might have.

The Facts

The first goal is simply to uncover facts, that is to say, the dates on which events occurred and the identities or location of its participants. For example, the date on which the German army invaded Poland is September 1, 1939. The evidence

for this fact includes eyewitness accounts, photographs, and contemporary newspaper reports of the events of the day.[4] The German army was not present on Polish soil on August 31 and was seen crossing the German-Polish border on September 1. The World Trade Center was attacked by hijacked planes, the first of which struck at 8:46 a.m. local time on September 11, 2001. Evidence for this fact also includes eyewitness accounts, photographs, and contemporary newspaper reports of the events of the day. Marie Antoinette was executed on October 16, 1793. This too is a fact. But other facts have yet to be uncovered, and it is perfectly reasonable for an academic historian to draw their university salary in their pursuit of them.

It is important to pause here and emphasize that the work of uncovering facts is neither straightforward nor easy, and many historical "facts" turn out to have been incorrect. Galileo never muttered "*eppur si muove* (yet it moves)" as he recanted his heliocentric beliefs (see, for example, Livio 2020); contra Hobsbawn, the number of troops deployed against the Luddites between 1811–13 did not "greatly exceeded in size the army which Wellington took into the Peninsula in 1808."[5] Alleged facts need to be checked. Moreover, facts of the past should not be confused with historical facts, which is to say, the facts a historian chooses to include in their account of what happened. "It is the historian who has decided for his own reasons that Caesar's crossing of that petty stream, the Rubicon, is a fact of history," writes E. H. Carr (1961), "whereas the crossing of the Rubicon by millions of other people before or since interests nobody at all … The belief in a hard core of historical facts existing objectively and independently of the interpretation of the historian is a preposterous fallacy, but one which it is very hard to eradicate." It is for this reason that a second goal of the historian should be considered.

The Account

Germany invaded Poland on September 1, 1939, and Polish forces counterattacked the German army at the Battle of the Bzura, which began on September 9. The attacks on 9/11 involved nineteen hijackers, some of whom had been trained at flight schools in the United States. The Austrian roots of Marie Antoinette made people wary of her loyalties. In giving an account of what happened, the historian must first choose which facts to include, and, no less importantly, which to exclude. Facts are, as Carr notes, the raw materials out of which a narrative account is constructed. But it is the historian "who decides to which facts to give the floor, and in what order or context" (Carr 1961, 5). (Historical fiction might include some of the same facts but makes no claim that an imagined dialogue is factual.)

Facts by themselves are not enough to describe what happened. The historian of the Second World War cannot simply rattle off the facts of the war, like who invaded whom and which battles were fought and where. These facts do not tell

the reader how to understand them, how one event led to another, and which were more decisive. For that, an account is needed, also called the story or the narrative. The historian can provide (one version) of that story. First the facts, then the story.

The Explanation

The historian first establishes the facts of what happened, distills them into the ones that they consider historically important, and then weaves them into a story. But there is a third step beyond the account of what happened. It is the story of why it happened, and that narrative is not only broader but also even more subjective. Why did Germany invade Poland on September 1, 1939? Was it the conditions of the Treaty of Versailles or the unstable political system in which Hitler was able to flourish? Why did terrorists attack the World Trade Center on September 11, 2001? Was it poverty in parts of the Arab world, the Russian invasion of Afghanistan that resulted in radicalization, or United States policy in the Middle East? The history of the 9/11 attacks is more than the date on which they occurred. It is the story of how and why, an explanation in which facts are necessary but not sufficient.

One might propose that the history of each of these events be open to replication, allowing others to review the evidence for the facts, the account, and the explanation. As noted, if facts cannot be agreed upon, historians cannot reliably proceed to an account. Facts may be incorrect, but even if they are assumed as stipulated, there may be debate as to which rise to the level of a historical fact.[6]

An Example from the History of Science and Religion

Like any author, the works I know best are the ones I have written. In giving an account of the Jewish reception of Copernican thought, I first had to establish facts of the past (Brown 2013). Which was the very first Jewish text to mention Copernicus by name? Which was the first to accept his heliocentric theory, and which the first to reject it? Until these and many other facts of the past had been established to my satisfaction, it was not possible to begin an account of the Jewish reception of Copernican thought. But in telling that story, which is to say, in giving an account of what happened, I had to make more subjective choices, not only about which facts of the past were important and therefore should be included but also the way to tell that story. Should it be based chronologically on what happened first and what happened next? Perhaps instead it should be arranged geographically, wherein the facts of the past are accounted for not exclusively based on when they occurred but also where. If another historian of science and religion were to tell the same story, would their choices be the same as mine? Could my account be replicated? This is a question that should address each part of my work.

I uncovered data, much of it previously ignored, that showed whether and when Jews (but not Judaism) accepted the Copernican model.[7] Did I get the dates correct? Did I overlook a fact or mistake a historical fact for a fact of history? Would another historian, given the same documentary facts, produce the same account? Would my explanation of why personalities took their particular pro- or anti-Copernican stances be the same as those of another? At each level, a degree of replication can be expected, and if my work could not be replicated, how else should this story and explanation be crafted?

For each part of the work of the historian, the replication effort will vary. Rather than question whether there can be a replication program in history (or literature, or art, or whatever), they need to be very explicit about what kind of history we are attempting to replicate. This should not be thought of as a weakness that applies to replication in the humanities and not in the sciences, because as I will discuss, replication in the sciences is only available for a limited subset of scientific experiments.

A Brief History of the Replication Crisis

The crisis of replication perhaps began around 2010, when two science journalists launched a website to publicize not only scientific papers that had been retracted but also the story of why (Collier 2011). Some of these retractions were based on fraudulent data, in which a published paper described a scientific experiment that may never have been performed or used data that may have been wholly or partially fabricated. But the results of a scientific experiment need not be based on fraudulent data to be misleading. They may be unlikely outcomes that occurred as a matter of chance but cannot (and this is important) be replicated.

Perhaps the most infamous example of this is the 2010 paper by Dana R. Carney, Amy J. C. Cuddy and Andy J. Yap on the effect of high-power poses ("expansive positions with open limbs") on testosterone and cortisol (two hormones associated with "dispositional and situational status and dominance and feelings of power"). Forty-two volunteers were randomly assigned to a high-power-pose or low-power-pose condition, and the results demonstrated, at least to the satisfaction of the authors, that high-power displays (as opposed to low-power displays) caused "physiological, psychological, and behavioral changes consistent with the literature on the effects of power on power holders—elevation of the dominance hormone testosterone, reduction of the stress hormone cortisol, and increases in behaviorally demonstrated risk tolerance and feelings of power" (Carney, Cuddy, and Yap 2010). One of the authors, Cuddy, went on to have a highly successful if somewhat brief career in the popular press as an expert on the use of power poses in industrial psychology. At the time of this writing, she had given the second most highly watched TED talk of all time, amassing over sixty-nine million views. And four years after the paper, she published *Presence: Bringing Your Boldest Self to Your Biggest Challenges*,

which was a *New York Times* bestseller, and one of the *Forbes* 15 Best Business Books of 2015.

There was just one problem. Almost no one could replicate this study. A 2015 paper published in the same journal as the original power-pose study reported that in experiments with 200 volunteers (a sample size some four times larger than used in the original study) there was no effect on hormonal levels or on any of three behavioral tasks (though it did effect self-reported feelings of power) (Ranehill et al. 2015). The authors of the original study responded with a common retort: the methods of the 2015 study were different. At least seven other attempts at replication failed to find any change in preregistered behavioral or hormonal outcomes (Loncar 2021; Jonas et al. 2017). Carney (2016), the first author of the original paper, has since disavowed its findings. "As evidence has come in over these past 2+ years," she wrote, "my views have updated to reflect the evidence. As such, I do not believe that 'power pose' effects are real."

The power-pose paper exposed a problem that is now of major concern: many results of experiments performed in the social sciences cannot be verified. In 2015, the Open Science Collaboration (2015) published its efforts to replicate 100 experimental and correlational studies that had been reported in three psychology journals. They found that replication, which may be variously defined, was limited. The mean effect size in the replication studies—a quantitative measure of the magnitude of the experimental finding—was only half of that seen in the original studies, and although 97% of the original studies reported statistically significant results ($p = < 0.05$), only 36% of the replicated studies did. This problem of replication has also been found in what might be called the hard sciences. "Across multiple criteria," the authors of one basic cancer biology study wrote, "the replications provided weaker evidence for the findings than the original papers" (Errington et al. 2021). For example, the median effect size for the replications was 85% smaller than the median of the original effect sizes. It has been difficult to replicate the success reported in many early drug trials for cancer. When, over a decade, the biotechnology giant Amgen tried to replicate the findings of fifty-three "landmark" hematology and oncology studies, they could do so in only six (11%) (Begley and Ellis 2012). So, a failure to successfully replicate, by which of course I mean to confirm the findings of earlier studies, is clearly not confined to the social sciences. Some 52% of those working in the basic sciences believe that there is a significant crisis of reproducibility, (although perhaps somewhat counterintuitively, 74% said they think at least half of the papers in their field can be trusted) (Baker 2016). Several other disciplines have noted a crisis in reproducibility, including chemistry (Bergman and Danheiser 2016), economics (Camerer et al. 2016), hydrology (Stagge et al. 2019), neuroanatomy (Marek et al. 2022; Poldrack et al. 2017), and human clinical trials (Van Noorden 2023).

What Do We Talk about When We Talk about Replication?

Any early attempts to verify results reported in the social sciences or basic sciences literature must first address an important question: What are we talking about when we talk about replication? What, precisely are the criteria for a successful replication? There are several candidates, and the ones which are chosen will determine whether or not the replication was successful.

Some might suggest using the effect size as the criteria for a successful replication. This is a measure of the magnitude of the difference and direction between the (means of the) two groups compared, and provides a context in which to consider the practical significance of the results. Indeed, for several years some of the major journals have required that the effect size (and not just the *p* value) be reported in all experimental papers (Durlak 2009). But even a goal as clear as this has another degree of complication, centering on the question of which statistical tests should be deployed. There are many statistical measures that might be legitimately used, including raw group differences, Cohen's *d*, which is used to measure the effect size of the means of two groups, and the odds ratio, which provides the odds of a successful outcome in the intervention group relative to the control group. Another measure could also be the direction of the outcome—whether or not there was, overall, a positive or negative finding that matched the direction found in the original study.

The decision about which tests, or group of tests, are to be used in a replication study is far from an objective process. Consider a case in which a replication study results in a finding in the same direction as the original but whose effect was much smaller. Is this to be considered to be a successful or unsuccessful replication? The 100 replication studies undertaken by the Open Science Collaboration used four quantitative markers of replication. But there was also a surprising question that had a yes–no answer and was itself a subjective measure: Did it replicate?

There are three further points I want to emphasize. First, there are some experiments, particularly in clinical science (where they are called trials), that are technically impossible to reproduce. Some clinical trials include many thousands of subjects across multiple countries, cost many millions of dollars, and take many years to perform. The use of energy, resources, and limited research funds to reproduce one's results, while a desideratum certainly of great value, is all but guaranteed not to happen.[8] Second, a failure to reproduce a scientific finding does not in itself indicate that there was misconduct (KNAW 2018). As the previous director the National Institutes of Health notes, a number of other factors may have contributed to a failure, including the poor training of researchers and prior publications that did not report important elements of experimental design, making their replication very challenging (Collins and Tabak 2014). Third, even if a study is successfully replicated, it should not automatically be concluded

that it was credible. "Successful replication increases confidence that the finding is repeatable," the group from the Center for the Open Sciences writes, "but it is mute to its meaning and validity" (Errington et al. 2021). They continue: "For example, if the finding is a result of unrecognized confounding influences or invalid measures, then the interpretation may be wrong even if it is easily replicated. Also, the interpretation of a finding may be much more general than is justified by the evidence. The particular experimental paradigm may elicit highly replicable findings, but also apply only to very specific circumstances that are much more circumscribed than the interpretation" (Errington et al. 2021).

Over many decades, those in drug development have learned that replication in different genders and ethnicities is vital. A new drug found to be successful in women may be less so in men (Soldin and Mattison 2009; Tamargo et al. 2017). Some drugs have an enhanced effect in certain ethnic groups compared to others (Chekka et al. 2021). But beyond these known differences, there are cases in which researchers have been oblivious to the profound effects of "unrecognized confounding influences" (Errington et al. 2021). If this is certain in the sciences, it is likely so in the humanities. This may be the greatest challenge to the replication program.

Degrees of Freedom

I now turn to what I believe is the fundamental challenge to replication in the humanities, and it comes from a statistical measure called degrees of freedom: the number of independent parameters that function within a system. Degrees of freedom may be thought of as the number of independent bits of information used to calculate a statistic, although the calculation can vary between statistical tests (Lazic 2010). Scientists can of course only account for degrees of freedom when they know about them; the challenge is that sometimes unknown or unrecognized confounding influences turn out to be additional degrees of freedom that were originally discounted.

Here is an example that, while necessarily detailed, is illustrative of this effect. Let us assume a researcher is interested in developing a new analgesic. To do so, the candidate drug must first be tested for safety and efficacy in animals. But how, precisely, can the analgesic ability of a drug on, say, a laboratory mouse be measured? Since it is known that humans grimace when in pain, it seems reasonable to assume that this response might be present in other mammals, including mice. The team develops a mouse grimacing scale that will give a reliable and reproducible measure of the subjective degree of pain felt by a mouse (Langford et al. 2010). The next step is to inflict pain on the mice, perhaps by injecting an irritant into their ankle joints. The laboratory technician may then observe the degree to which the mice grimace from the pain with and without the experimental analgesic.

These were the very steps taken by a team of researchers led by Robert Sorge from McGill University (Sorge et al. 2014). However, as they proceeded, they noticed that their experimental mice failed to grimace in ways that would be expected. But this changed when the laboratory technician left the room. Once alone, the mice began to exhibit their usual behaviors, including grimacing, associated with pain. Somehow, the presence of the technician was affecting the behavior of the mice; it appeared to provide them with a degree of analgesia. But this was (quite literally) only half the story, because it only happened when the technician was male. The presence of female technicians did not prevent the usual grimacing. Eventually, the team recognized that the mice were responding to the smell rather than the presence of a male technician, because the same grimace repression occurred in the presence of clothing that had been worn by a male. In fact, the team learned that the smell of any number of males from different species could prevent grimacing: the mice reacted to the presence of the smells from male guinea pigs, male cats, and male dogs in the same way they reacted to the presence of a male laboratory technician.[9]

I have detailed this episode because it highlights the challenge of reproducibility and the number of degrees of freedom. Before this gender-based finding was observed, were a team to reproduce the original experiments by Sorge and his colleagues, they might have used the same strains and gender of mice and the same technique for inducing pain. But would they have performed the experiment at the same times of the day or during the same season as the original? Would they have used a technician who was the same gender? If these were thought to be extraneous factors that would not alter the outcomes of the experiment, then these independent parameters, these degrees of freedom, would not have been replicated. If the original experiment had been performed on a rainy day when the barometric pressure was low, it could reasonably be repeated on a sunny day when the pressure is high, if, and only if, this degree of freedom (the barometric pressure) was thought to be inconsequential. But it turns out that even in the controlled environment of a laboratory setting, some discounted degrees of freedom can unexpectedly become critically important. And one of these that had been previously ignored was the gender of the technician. The effects of the gender of the experimenter have also been found in different human studies. (I am cognizant of the irony of citing the following experiments, which may not yet have been replicated.) People seem to perform better on memory tests when the experimenter is of the opposite sex; they may also have better physical performance when the experimenter is of the opposite gender. Men have elevated testosterone when the experimenter is a woman and seem to tolerate more pain than if the experimenter were male (Chapman, Benedict, and Schiöth 2018). This previously unrecognized confounding influence has introduced a new degree of freedom into the experimental method.

Are there limits to what degrees of freedom are allowed for? It is always assumed that there are, because some must be discounted to avoid a paralysis of investigation. For example, when replicating experimental conditions in the basic or social sciences, the astrological configuration is not controlled for; the experimenter is not required to perform their test when Jupiter is in the constellation of Aires (or whatever) because those were the astrological conditions under which the original experiment was performed. Astrology is discounted as a degree of freedom, just as the gender of the experimenter once was. Might we change our minds about astrology? Of course. If the evidence demonstrated that astrological conditions influenced the behavior of subjects, be they human or non-human animals or cells, then it would be correct to include this variable as another degree of freedom, just as the gender of the experimenter is now included in some rodent experiments. So far, of course, there is no such evidence, and so astrology is correctly discounted as not being of any importance, of adding a degree of freedom, in the replication of previous work (Thagard 1978; Moberger 2020).

The Encroachment of Scientism

If a program of replication is achievable for (at least some parts of) the humanities, what becomes of the road towards scientism, the belief that the world is more explainable with the help of basic scientific principles? Might its claims be strengthened? The minimalist or weak version claims that the methods of science are the best ways of securing knowledge of anything, and the maximalist or strong version claims that the methods of science are the only reliable ways (van Woudenberg 2023). But both versions might encourage those in the humanities to focus only on those disciplines that allow for some kind of replication. Only they would be considered worthy of academic study. It is not hard to imagine a scenario in which professors of literature or history would be expected to support their salaries and PhD students with research funding, just as is now the case with academics who study and teach the basic sciences (Boss and Eckert 2004). Only the humanities—or better, only parts of the humanities—that allow for replication would be supported. Work outside this niche would no longer be welcome in the fellowship of higher education.

This danger must be addressed by those who would advocate for an agenda of replication in the humanities. University undergraduates are already leaving the liberal arts for degrees that focus on health, technology, and business, and this trend will only worsen if scientism ascends.[10] Leon Wieseltier (2013) has warned that as a result of scientism, "the humanities are the handmaiden of the sciences, and dependent upon the sciences for their advance and even their survival," and the replication agenda may well increase this dependence, in

terms of both methodology and financial support. Beware of the unintended consequences that flow from a stance of confidence in the humanities.

Putting It All Together

In one of the first papers on this topic, Peels (2019) describes four potential obstacles to replication in the humanities. First, that the object or event under study is unique. Second, the methodologies employed in the humanities do not lend themselves to replication. Third, many of the objects of study in the humanities "are normative in the sense that they are objects of value and meaning, whereas this is not the case in many of the natural and biomedical sciences." And fourth, even though replication may well be possible in the humanities, "it is not particularly desirable—not something to aim at or invest research money on—because there is simply too much disagreement in the humanities for there to be a successful replication sufficiently often." To each of these obstacles, Peels offered a rejoinder: few events are really unique (there was only one French Revolution, but there were many other revolutions); many different kinds of methodologies are used in the social sciences, and some (but not all) may indeed be replicable (those that are empirical but not those methods that are deductive); even when the humanities are concerned with meaning (and the sciences with molecules,) it may be possible to replicate that meaning multiple times; and finally, disagreement, or better, differing schools of thought, are a feature of many disciplines outside of the humanities, such as economics and quantum physics, and yet these disciplines may still be subject to replication.

It is now clear that while Peels believes that replication in the humanities may be possible but only in a limited and circumscribed way, this should not be thought of as a constraint that applies only to the humanities. As I have carefully noted, replication in a limited and circumscribed way is also a feature of basic and clinical research. Some degrees of freedom are replicated, while others, thought not to be important, are not. It will take time and further effort to determine whether these were correct. But at its core, this article has demonstrated that there can be a replication program in some of the humanities and that a constricted version of replication does not *ab initio* suggest that the program is of little value.

Let us return to the three kinds of history and see how they are each affected by the choice of degrees of freedom. First, I noted that historians choose from a pool of facts that they decide are to be considered historical. These facts are subject to replication, or better, verification. But the choice of facts, the decision to declare Caesar's crossing of the Rubicon, but not that of millions of other people before or since, a fact of history is far more subjective and falls outside of a replication program. In addition, there may be historical facts that had a previously unnoticed effect. Once discovered, they, like the gender of the laboratory technician, may no longer be discounted.

Second, the decision to develop the chosen historical facts into an account may also be subject to replication. But the danger from unknown degrees of freedom increases as ever more expansive accounts are developed and as historians move from the facts to the account and then to the explanation. One historian may include previously discounted degrees of freedom and produce a historically coherent account that is at odds with another's. When challenged, each may simply explain that their account includes degrees of freedom that were deliberately or accidentally discounted by the other. This is particularly obvious when it comes to the third kind of history told: the explanation of what happened. One historian believes Napoleon's doomed march on Moscow failed because of the poor logistics and a failure to supply his forward forces. A second claims that another degree of freedom needs to be accounted for: the weather. For a third, there is a new degree of freedom: the altitude of the terrain. A fourth claims that the pivotal degree of freedom, one that had been overlooked (and so was an unrecognized confounding influence), was the failure of Napoleon's subordinates to properly command and control their forces (Keefe 2015). As the degrees of freedom expand, the likelihood for any meaningful replication diminishes. But, as demonstrated by the gender of lab technicians, this is also true of the sciences.

Ever since Karl Popper, most believe that a scientific statement is one that at its basis is subject to falsification.[11] It only takes one black swan to disprove the thesis (which in this case is nothing more than an observation) that all swans are white. The concept of falsifiability is to empirically test a theory or hypothesis to see if it is false. Replication is the accepted way of verifying this. But let us remember that not all hypotheses or observations are scientific, in so much as they cannot be falsified. But this does not make them less important. $E = mc^2$ is a scientific explanation; proclaiming that "I love you" is not. But both are enriching. What then, is to be done with non-falsifiable theories in the humanities, those that do not lend themselves to verification? Discard them as unscientific, or embrace them because they too enhance our lives? A scientific account of the pigments that make up the eyes of Johannes Vermeer's *Girl with a Pearl Earring* will do nothing to explain why her gaze is so haunting. For that, the non-verifiable and non-falsifiable explanations of emotion, empathy, and passion are still needed.

Acknowledgments

The views expressed are solely those of the author and do not represent the official views of any branch of the United States federal government.

Notes

[1] The Second Amendment is brief: "A well regulated Militia, being necessary to the security of a free State, the right of the people to keep and bear Arms, shall not be infringed."

[2] This is not the place for a discussion of textualism versus originalism.

[3] See the articles in this thematic section, particularly the Introduction (Peels et al.), the conceptual replication (Pear, van den Brink, and Peels), and the article by Brooke.

[4] See, for example, the front page of London's *Evening Standard* newspaper on September 1, 1939, available at https://time.com/5659728/poland-1939/.

[5] Hobsbawn first made this claim in a paper published in 1952 and repeated it in his book *Labouring Men,* published in 1964 (see Hobsbawn 1952, 58; 1965, 6). For an analysis, see Kevin Linch (2011, 4–5).

[6] These ways of "doing" history are of course not meant to be exhaustive. Understanding what an event meant to the people who lived through it is another, and no less important, way of writing history.

[7] Since the destruction of the Second Temple and the dissolution of the Sanhedrin, there has been no central authority for all of Jewish practice. In eastern Europe for example, most towns had their own rabbinic authority and rabbinic court.

[8] It is for this reason that it is so important to get it right the first time.

[9] On reflection, this finding should not have been surprising. Mouse olfaction is highly developed; the scent of mouse urine transmits information about the individual identity of its owner, as well as its reproductive status, health, and food resources (see Hurst and Beynon 2004, 1288–98).

[10] In the United States, there has been an enormous rise in the numbers of STEM undergraduate degrees: over the past ten years, there has been a doubling of the percentage of bachelor's degrees awarded in computer and information science, a 32% increase in mathematics and statistics degrees, and a 47% increase in undergraduate degrees in engineering. These come at the expense of the liberal arts. Over the same period, there was a 10% decrease in the number of bachelor's degrees awarded in social sciences and history, a 16% decrease in philosophy and religious studies degrees, and a 32% decrease in English degrees. Data extracted from National Center for Education Statistics (2023).

[11] Most, but not all. There are other accounts of what make a statement "scientific," but for the sake of space, only the Popperian definition is addressed. For a survey of the other definitions see, for example, https://plato.stanford.edu/entries/pseudo-science.

References

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604): 452–54.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–33.

Bergman, Robert G., and Rick L. Danheiser. 2016. "Reproducibility in Chemical Research." *Angewandte Chemie International Edition* 55 (41): 12548–49.

Boorstin, Daniel. 1989. *Hidden History.* New York: Vintage Books.

Boss, Jeremy M., and Susan H. Eckert. 2004. "Academic Scientists at Work: Giving It 110%." *Science*, February 13, 2004. https://www.science.org/content/article/academic-scientists-work-giving-it-110.

Brown, Jeremy. 2013. *New Heavens and a New Earth: The Jewish Reception of Copernican Thought*. New York: Oxford University Press.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Carney, Dana R. 2016. "My Position on Power Poses." https://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf.

Carney, Dana R., Amy J. C. Cuddy, and Andy J. Yap. 2010. "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21 (10): 1363–68.

Carr, Edward Hallet. 1961. *What Is History?* London: Macmillan and Co.

Chapman, Colin D., Christian Benedict, and Helgi B. Schiöth. 2018. "Experimenter Gender and Replicability in Science." *Science Advances* 4 (1): e1701427.

Chekka, Lakshmi Manasa S., Arlene B. Chapman, John G. Gums, Rhonda M. Cooper-DeHoff, and Julie A. Johnson. 2021. "Race-Specific Comparisons of Antihypertensive and Metabolic Effects of Hydrochlorothiazide and Chlorthalidone." *American Journal of Medicine* 134 (7): 918–25.

Collier, Roger. 2011. "Shedding Light on Retractions." *Canadian Medical Association Journal* 183 (7): E385–86.

Collins, Francis S., and Lawrence A. Tabak. 2014. "Policy: NIH Plans to Enhance Reproducibility." *Nature* 505 (7485): 612–13.

Durlak, Joseph A. 2009. "How to Select, Calculate, and Interpret Effect Sizes." *Journal of Pediatric Psychology* 34 (9): 917–28.

Errington, Timothy M., Maya Mathur, Courtney K. Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A. Nosek. 2021. "Investigating the Replicability of Preclinical Cancer Biology." *eLife* 10:e71601.

Hobsbawn, Eric. 1952. "The Machine Breakers." *Past & Present* February:57–70.

Hobsbawn, Eric. 1965. *Labouring Men: Studies in the History of Labour*. London: Weidenfield and Nicolson.

Hurst, J. L., and R. J. Beynon. 2004. "Scent Wars: The Chemobiology of Competitive Signalling in Mice." *Bioessays* 26 (12): 1288–98.

Jonas, Kai J., Joseph Cesario, Madeliene Alger, April H. Bailey, Dario Bombari, Dana Carney, John F. Dovidio, et al. 2017. "Power Poses: Where Do We Stand?" *Comprehensive Results in Social Psychology* 2 (1): 139–41.

Keefe, John M. 2015. *Failure in Independent Tactical Command: Napoleon's Marshals in 1813*. n.p.: Pickle Partners Publishing.

KNAW (Royal Netherlands Academy of Arts and Sciences). 2018. *Replication Studies: Improving Reproducibility in the Empirical Sciences*. KNAW: Amsterdam.

Langford, Dale J., Andrea L. Bailey, Mona Lisa Chanda, Sarah E. Clarke, Tanya E. Drummond, Stephanie Echols, Sarah Glick, et al. 2010. "Coding of Facial Expressions of Pain in the Laboratory Mouse." *Nature Methods* 7 (6): 447–49.

Lazic, S. E. 2010. "The Problem of Pseudoreplication in Neuroscientific Studies: Is It Affecting Your Analysis?" *BMC Neuroscience* 11:5.

Linch, Kevin. 2011. *Britain and Wellington's Army: Recruitment, Society and Tradition, 1807–15*. Basingstoke, UK: Palgrave Macmillan.

Livio, Mario. 2020. "Did Galileo Truly Say, 'And Yet It Moves'? A Modern Detective Story." *Scientific American*, May 6, 2020.

Loncar, Tom. 2021. "A Decade of Power Posing: Where Do We Stand?" *Psychologist* 34:40–45.

Marek, Scott, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, et al. 2022. "Reproducible Brain-Wide Association Studies Require Thousands of Individuals." *Nature* 603 (7902): 654–60.

Moberger, Victor. 2020. "Bullshit, Pseudoscience and Pseudophilosophy." *Theoria* 86 (5): 595–611.

National Center for Education Statistics. 2023. "Undergraduate Degree Fields: Condition of Education." https://nces.ed.gov/programs/coe/indicator/cta.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Peels, Rik. 2019. "Replicability and Replication in the Humanities." *Research Integrity and Peer Review* 4 (1).

Peels, Rik, and Lex Bouter. 2018. "The Possibility and Desirability of Replication in the Humanities." *Palgrave Communications* 4 (1): 95.

Poldrack, Russell A., Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, et al. 2017. "Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research." *Nature Reviews Neuroscience* 18 (2): 115–26.

Ranehill, Eva, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. 2015. "Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women." *Psychological Science* 26 (5): 653–56.

Schlesinger, Arthur M. 1957. *The Colonial Merchants and the American Revolution, 1763–1776*. New York: Frederick Ungar Publishing.

Soldin, Offie P., and Donald R. Mattison. 2009. "Sex Differences in Pharmacokinetics and Pharmacodynamics." *Clinical Pharmacokinetics* 48 (3): 143–57.

Sorge, Robert E., Loren J. Martin, Kelsey A. Isbester, Susana G. Sotocinal, Sarah Rosen, Alexander H. Tuttle, Jeffrey S. Wieskopf, et al. 2014. "Olfactory Exposure to Males, Including Men, Causes Stress and Related Analgesia in Rodents." *Nature Methods* 11 (6): 629–32.

Stagge, James H., David E. Rosenberg, Adel M. Abdallah, Hadia Akbar, Nour A. Attallah, and Ryan James. 2019. "Assessing Data Availability and Research Reproducibility in Hydrology and Water Resources." *Scientific Data* 6:190030. https://doi.org/10.1038/sdata.2019.30.

Tamargo, J., G. Rosano, T. Walther, J. Duarte, A. Niessner, J. C. Kaski, C. Ceconi, et al. 2017. "Gender Differences in the Effects of Cardiovascular Drugs." *European Heart Journal: Cardiovascular Pharmacotherapy* 3 (3): 163–82. https://doi.org/10.1093/ehjcvp/pvw043.

Thagard, Paul R. 1978. "Why Astrology Is a Pseudoscience." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1978 (1): 223–34. https://doi.org/10.1086/psaprocbienmeetp.1978.978227.

Van Noorden, Richard. 2023. "Medicine Is Plagued by Untrustworthy Clinical Trials. How Many Studies Are Faked or Flawed?" *Nature* 619 (7970): 454–58. https://doi.org/10.1038/d41586-023-02098-7.

van Woudenberg, René. 2023. "Argumentative Strategies against Scientism: An Overview." *Interdisciplinary Science Reviews* 48 (2): 1–16.

Wieseltier, Leon. 2013. "Crimes against Humanities: Now Science Wants to Invade the Liberal Arts. Don't Let It Happen." *The New Republic*, September 16, 2013.