Not-So-Natural Experiments in History*

Christian Dippel† Bryan Leonard‡

December 14, 2020

Abstract

The paper compares the role of cliometrics — broadly defined to include economics, political science, and other social sciences — before and after the "credibility revolution" of the late 1990s. The contributions of cliometrics that led to the 1993 Nobel Prize were due primarily to a combination of quantification and economic theory with in-depth historical knowledge. After the credibility revolution, much of cliometrics shifted towards "natural experiments," especially in papers published in general-interest journals. We argue that this shift comes with certain trade-offs between statistical and contextual evidence, and that the refereeing process currently makes these trade-offs steeper in historical settings than in other observational-data settings. We also argue, however, that historical settings offer particularly actionable ways of flattening these trade-offs to ensure the "clio" in cliometrics stays alive and well.

*For helpful comments we thank Jeff Jenkins, one anonymous referee, Doug Allen, Terry Anderson, Lee Alston, Eric Edwards, P.J. Hill, Gary Libecap, and Steven Smith.

†University of California, Los Angeles, and NBER

‡Arizona State University

"History matters...Today's and tomorrow's choices are shaped by the past and the past can only be made intelligible as a story of institutional evolution. Integrating institutions into economic theory and economic history is an essential step in improving that theory and history."
Douglass C. North, *Institutions, Institutional Change, and Economic Performance*

1 Introduction

The field of cliometrics began in 1957 with a joint conference held by the Economic History Association and the National Bureau of Economic Research (North, 1977). Over 30 years later, Robert Fogel and Douglass North won the 1993 Nobel Prize[1] in economics for "having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change." The Nobel committee recognized Fogel and North as "pioneers in the branch of economic history that has been called the 'new economic history,' or cliometrics, i.e. research that combines economic theory, quantitative methods, hypothesis testing, counterfactual alternatives and traditional techniques of economic history, to explain economic growth and decline" (Royal Swedish Academy of Sciences, 1993). In their survey of the "cliometrics revolution," Lyons, Cain, and Williamson (2007) characterize four distinctive pillars of cliometrics: *(i)* the use of quantifiable evidence, *(ii)* the use of theoretical concepts and models, *(iii)* the use of statistical inference, and *(iv)* the use of a historian's skill to both judge source material and place it in institutional and historical context.[2] All four components are evident in the description of the Fogel/North Nobel prize.

In the nearly three decades since the Nobel there have been two major changes in cliometrics as it kept pace with broader changes in empirical methods in economics and political science. First, whereas Fogel and North won the Nobel in part because their work "not only increased our knowledge of the past, but has also contributed to the elimination of irrelevant theories" (Royal Swedish Academy of Sciences, 1993), modern cliometric endeavors are most heavily scrutinized through the lens of statistical identification and research design. Second, and relatedly, there has been an increasing trend of publishing economic history papers in general-interest journals

[1] Technically, the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel.
[2] We follow Lyons et al. (2007) in taking a broad definition of cliometrics that encompasses not just economics, but also political science and other social sciences.

(Abramitzky, 2015; Margo, 2018). These two changes were facilitated by dramatic improvements in the ease of archival data collection, computing power, and the widespread uptake of spatial data for use in econometric analysis. More than anything, however, these changes were driven by the "credibility revolution" that swept economics from the mid-1990s and placed heavier emphasis on distinguishing causal relationships from partial correlations using quasi-experimental techniques such as instrumental variables, regression discontinuity, and panel fixed effects (Angrist and Pischke, 2010).

Cliometrics has no doubt benefited from these advances, but not without creating trade-offs. As the profession has put increasing weight on empirical methods, techniques such as balance tests and placebo checks have partially substituted for qualitative contextual arguments as tools for making the case for the validity of a quasi-experimental design (Diamond and Robinson, 2010). Along the way, this emphasis on quasi-experimental causal identification strategies has created a benefit to simplifying historical context to highlight a single "general interest" mechanism, and to de-emphasizing historical context that muddies the waters around this mechanism. In combination, these factors have given the modern researcher an incentive to *downplay* historical context because it is costly for referees to adjudicate and potentially complicates the presentation of a clean empirical design.

This incentive is of course also present when researchers study policies in contemporary settings. However, the problem is likely to be more pronounced in cliometrics because in contemporary settings a larger set of economists works on an inherently smaller range of settings. A basic knowledge of the underlying conditions that led to the passage of a policy and imperfections in how it was implemented can often be taken for granted in contemporary settings because the researcher (and the referee) inhabit that context themselves or are at least familiar with it. Take, for instance, the multitude of studies around the Clean Air Act.[3] Most environmental economists are familiar with the context surrounding to this act and the much-studied 1990 amendments to it, and almost all will have read around a half-dozen studies on the topic in graduate school. By contrast, the overwhelming majority of historical studies consider settings that are unfamiliar to the vast

[3]See, for example, Schennach (2000); Greenstone (2002, 2003); Chay and Greenstone (2005); Busse and Keohane (2007); Lange and Bellas (2007); Auffhammer, Bento, and Lowe (2009, 2011); Hubbell, Crume, Evarts, and Cohen (2010); Chan, Stavins, Stowe, Sweeney, et al. (2012); Walker (2013); Isen, Rossin-Slater, and Walker (2017); Bento, Freedman, and Lang (2015)

majority of cliometricians. Referees will therefore often read papers without prior knowledge that may be critical to a paper's identification strategy. These factors make it more difficult to detect differences between a simplistic vs. realistic reading of historical policy, and create an asymmetry of contextual knowledge between authors and referees. This problem is most pronounced when the natural experiment originates in a policy or an institution (as opposed to, say, geographic or climatic variation) because these are the settings where a referee without context-specific historical knowledge is at the biggest disadvantage. We therefore argue that Lyons et al.'s pillar *(iv)* needs to be strengthened again to validate the historical and contextual assumptions underlying modern empirical designs, especially when the natural experiments are rooted in policies or institutions, e.g., "the humanly devised constraints that structure political, economic, and social interaction" (North, 1991).

Fortunately, cliometrics is perhaps uniquely suited to overcome the problem of asymmetric contextual knowledge between authors and referees. Whereas the relevant subject-matter experts for policies studied in other fields may often be non-academic practitioners whose input can be difficult to solicit, in cliometrics there is often an abundance of relevant subject matter experts in the field of history. Even if they are untrained in causal identification, historians are familiar with the norms in academic publishing and possess considerable contextual knowledge that could be used to validate the plausibility of a natural experiment.

Indeed, cliometricians have traditionally engaged actively with historians and held each other to a high standard within the field (Margo, 2018). Today, however, this is less true. Economic history as it is practiced in history departments has moved in the opposite direction of economic history as it is practiced in economics and political science departments, away from the use of quantitative evidence and generalization towards culturally grounded, context-specific analysis (Diamond and Robinson, 2010). As a result, it is today entirely possible for an economic history paper to be submitted to a general-interest journal without ever having been scrutinized by an audience with the relevant historical expertise.

Furthermore, a lack of scrutiny over a paper's historical plausibility may not be compensated in the peer-review process. In general-interest journals in particular, the peer-review process will often select referees who are at a disadvantage for evaluating the plausibility of so-called natural experiments drawn from history because the editor may have difficulty finding a referee with the

relevant contextual knowledge (especially in the typical case where the editor themselves is not an economic historian), or because they may not want to "use up" a full referee report on a historian who may be unable to comment on large parts of the paper devoted to statistical identification. This potential problem is illustrated succinctly in an episode recalled by Gavin Wright where he received a referee report stating that a specific "paper might be acceptable in the *AER*, but it does not meet the standards of the *JEH*" (Abramitzky, 2015).

We therefore propose an addition to the standard peer review process: authors submitting a cliometric paper to a general-interest (or any) journal should include a separate "historical abstract" that delineates the purely historical parts of the paper and that clearly states which parts of the history give rise to the natural experiment. If the set of experts with the relevant historical knowledge are unwilling or unable to provide a full report commenting on the parts of the paper devoted to statistical identification, the editor could instead solicit narrow and brief feedback from at least one expert historian that only comments on the veracity and plausibility of the historical natural experiment as it is described in the paper.

2 The Historical Contributions of Cliometrics

The contributions of cliometrics that led to the 1993 Nobel Prize leveraged the tools of economics in pursuing quantification, theory, and classical statistical inference (Lyons et al.'s pillars *(i)–(iii)*). We argue that they did so while maintaining the "historian's skill" of putting a strong emphasis on institutional and historical context (Lyons et al.'s pillar *(iv)*). This section provides a brief discussion of the prominent role of historical knowledge in cliometrics prior to the credibility revolution. In Section 3, we juxtapose this with the emergence of quasi-experimental methods in cliometrics after the 1990s, whose application has, we argue, on occasion come at the cost of paying insufficient attention to historical context.

2.1 Quantifying Historical Phenomena

Before the rise of cliometrics, historians were often criticized for making quantitative claims (e.g. "X decreased over time," or "A was more significant than B," or "C led to a rise in D") without providing the numbers to back up these claims (Diamond and Robinson, 2010). Quantitative

historians have therefore increasingly been expected to substantiate their assertions about quantitative relationships using statistical or at least numerical evidence (Stone, 1979). Quantification was thus the hallmark of the early cliometricians who improved the quantitative study of history and enabled the evaluation of quantitative claims by digitizing and tabulating historical data, record-linking historical data-sets, and cleverly measuring difficult-to-quantify concepts (North, 1977; Eichengreen, 1994; Margo, 2018).

In addition to substantiating (or disproving) inherently quantitative claims, cliometricians have provided an important counterpoint to other forms of evidence relied upon by historians. Historians have for instance often focused on first-hand accounts and other contemporary descriptions of events to construct their understanding of historical episodes (Stone, 1979). While rich in context, these accounts can be misleading, biased, or ambiguous and fail to provide an accurate picture of events as they happened. Take U.S. newspapers as an example. Because they were openly partisan for most of the Nineteenth Century, a historical study of the Reconstruction period that draws exclusively on Democratic-leaning newspapers would end up characterizing the motivations and merits of Reconstruction in a dramatically different way from one that draws exclusively on Republican-leaning newspapers.[4] Of course, this is an example where the bias is particularly obvious, and a well-trained historian is unlikely to make the mistake of exclusively relying on one set of partisan news coverage, but the broader point is that cliometrics can help enhance our understanding of history by quantifying the claims in question to adjudicate conflicting textual accounts.

An excellent example of the way in which cliometrics can complement text-based history on a topic is the debate about the origins of the U.S. constitution. This debate surrounds the question of why, 11 years after the Declaration of Independence, the founders decided to unite the 13 independent states back together in a country. Historically, the received wisdom had focused on the founders' high-minded ideals, immortalized in the Federalist Papers that preceded Philadelphia's constitutional convention. In 1913, however, Charles Beard introduced a powerful economic argument that the framers' core interest in writing the constitution was to get a federal government

[4] We seem to have recently re-entered such a world of overtly partisan newspapers. Future historians studying the 2020 presidential election drawing exclusively on coverage from the *New York Times*, *Washington Post* and *Daily Beast* would see a radically different picture than if they drew exclusively on the *New York Post*, *Washington Times* and *Daily Caller*.

that would enforce creditors' rights where the state legislatures of the 13 independent states had not. A debate then unfolded, in which most historians discounted Beard's narrative. The main counter-argument to Beard's argument was that the framers did not own substantial amounts of financial instruments that were being devalued by states' fiscal profligacy. While this counter-argument was inherently empirical, it was actually never well measured. Yet, it remained the received wisdom for many more decades until the mid-1980s, when McGuire and Ohsfeldt (1989) actually collected all the data on the state delegates' personal slave and debt holdings to undertake a careful cliometric analysis of Beard's argument. They documented that in addition to state delegates' ideology (the pre-Beard argument), their economic interests were in fact also highly predictive of their voting on constitutional ratification, in a manner that was highly consistent with that outlined by Beard.

2.2 Using Economic Theory in History

Quantification was not the only contribution of cliometrics to history. Lord Kelvin famously said that "when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind," to which Jacob Viner said "yes, and when you *can* express it in numbers, your knowledge is of a meager and unsatisfactory kind." Viner's statement was not a criticism of Kelvin, who was a physicist and engineer and as such operated in the hard sciences. Rather, it was a note of caution on the limits to quantification in the *social* sciences when it is not guided by theory (or—the topic of Section 3 —a claim to causal identification). Resonating with Viner's sentiment, Fogel, North, and other cliometricians married economic theory to data, statistical methods, and history, and thereby created a "new economic history" to replace the old, largely qualitative one (McCloskey, 1985; Eichengreen, 1994).

North's study of the economic development of the United States (1961) and Fogel's work on railroads and U.S. growth (1964) are classics in this regard. This new economic history combined all four of Lyons et al.'s pillars, and it fairly defined what cliometrics was between the mid-1960s and the mid-1990s.[5] It was for this that Robert Fogel and Douglass North won the Nobel in 1993.

Both Fogel and North argued that much of the old economic history was built on implicit theo-

[5]However, the "new economic historians" were constantly criticized for their new "ahistorical" approach by more traditional economic historians who were suspicious of the abstractions associated with the use of statistical hypothesis testing (North, 1977).

ries and they emphasized the value of making these theories explicit so that internal contradictions could be revealed and hypotheses could be tested (Fogel, 1967; North, 1977).[6] One way Fogel did this was by constructing explicit rather than implicit counterfactuals in his study of railroads in the United States, arguing that railroads generated modest savings relative to alternative modes of transportation Fogel (1964). In their study of slavery, Fogel and Engerman (1974) focused on comparing total factor productivity on plantations vs. in free agriculture, thus transforming "the debate from one predicated on the irrationality of the slave owner to one about the profitability of alternative economic systems" (Eichengreen, 1994).

The contributions above directly tied economic theory to the econometric analysis of historical forces. There is a distinct body of literature referred to as "analytical narrative approach" that also applies economic theory to shed light on history, but stands in contrast in that it does so without the use of econometrics. Greif's 1993 analysis of the Maghribi Trade networks was one of the earliest papers in this tradition, as was Greif, Milgrom, and Weingast's analysis of merchant guilds (1994). Other prominent examples of this tradition include, inter alia, Alston and Ferrie (1993), Puga and Trefler (2014), and Allen and Leeson (2015). This approach to economic history combines Lyons et al.'s pillars *(ii)* and *(iv)*, and possibly *(i)*, but generally not *(iii)*.

2.3 Cliometrics Since the 1993 Nobel

Quantification has continued to play a significant role in cliometrics since 1993, partly because the cost of large-scale digitization of archival materials has dramatically decreased with the scanning of millions of archival records, the emergence of professional data-entry firms in developing countries, and the development of new software and techniques for record linking (Abramitzky, 2015). Likewise, the application of economic theory to historical questions has continued to play a significant role after 1993. For example, Voigtländer and Voth (2013) recently provided an explanation of the European Marriage Pattern that was fully grounded in economic theory, and then tested their theory using newly collected data.[7]

[6] North's use of theory to make sense of history followed a somewhat different trajectory. Although his early contributions involved using neoclassical theory to explain economic growth (North, 1958, 1961), he became increasingly dissatisfied with the ability of neoclassical economics to account for the role of institutions in economic growth. This led him to focus more on building a theory of institutional change later in his career, as exemplified in North (1981, 1990, 2005)

[7] Voigtländer and Voth (2013) have inspired some healthy debate among historians. See, e.g. Dennison and Ogilvie (2014).

Cliometricians are also making use of theory-driven structural econometric approaches to answer new questions and shed new light on old ones (Jaworski, 2020). Gentzkow and Shapiro (2010) use a structural general-equilibrium model of newspaper demand to analyze the drivers of political "slant" among historical newspapers. Donaldson and Hornbeck (2016) revisit Fogel's 1964 examination of railroads' contribution to American economic growth with a structural approach for estimating changes in "market access" due to the development of railroads, an approach that has subsequently been adapted by Jaworski and Kitchens (2019) to study the impact of the Appalachian Highway Development System.

Despite this continuity, there has nevertheless been a marked overall shift away from theory in economic history. In large part, this is because cliometrics — in lock-step with applied econometric analysis more generally — has expanded into asking ever broader questions, many of which were previously confined to other fields of social science. One example of this is the recent evaluation of Turner's Frontier Hypothesis by Bazzi, Fiszbein, and Gebresilasse (2020). This paper developed novel ways of quantifying both the frontier and "individualism"' to address a long-standing historical debate about the evolution of culture that would traditionally have been outside the scope of standard economic analysis. Precisely because economic theory does not offer an analytic framework for what "frontier spirit" is, the authors were all the more reliant on sources outside economics to confirm the intuition of their approach. Hence, this study provides a nice example of how authors can engage with related historical literature to validate findings.

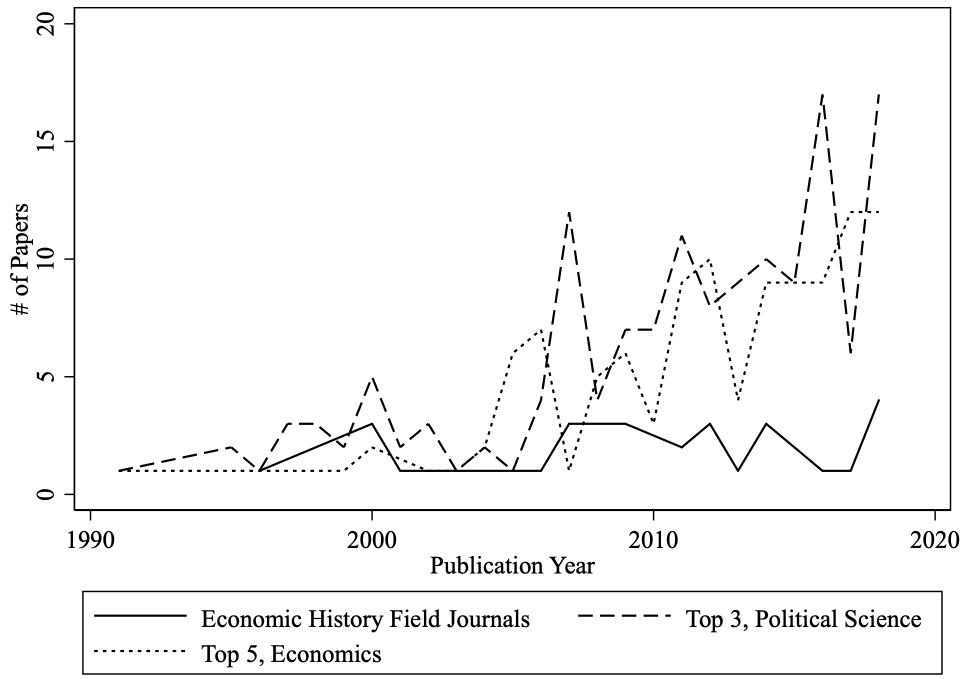
In an example from political science, Spirling (2012) uses systematic textual analysis to analyze whether outcomes were worse for Native American tribes whose treaties with the US government had harsher terms. In a sense, this shift away from theory was caused by the rise of quasi-experimental applied econometrics because causal identification has become a substitute to economic theory for qualifying a paper for a general-interest journal (Abramitzky, 2015). The next section describes this rise in quasi-experimental methods in cliometrics.

3 Natural Experiments in Cliometrics

The most recent contribution of cliometrics to the study of history arose out of the development of new methods in economics to infer causality from observational (e.g., historical) data (Angrist

and Pischke, 2010). The credibility revolution arrived on economic history's doorstep in the late 1990s; starting in particular with persistence studies that relate historical episodes to modern-day outcomes, often using historical shocks to get exogenous variation in present-day institutional settings (Nunn, 2009). Seminal early contributions include La Porta, Lopez-de Silanes, Shleifer, and Vishny (1997), Hall and Jones (1999), and Acemoglu, Johnson, and Robinson (2001). In the two decades since, natural experiments have become the gold standard in cliometrics (Abramitzky, 2015). In their 2010 volume *Natural Experiments of History*, Diamond and Robinson provide a useful working definition of natural experiments, which "consist of comparing...different systems that are similar in many respects but that differ with respect to the factors whose influence one wishes to study" (Diamond and Robinson, 2010).

Figure 1: History Papers Utilizing "Natural Experiments"



Notes: This figure depicts the number of papers focused on historical episodes that are characterized by their authors' as natural experiments over 1990 – 2019 in three sets of journals: the top five economics journals, the top three political science journals, and top field journals in economic history.

Figure 1 depicts the number of papers that are focused on historical episodes and that include the phrase "natural experiment," "policy experiment," or "quasi-experiment" over 1990–2019 in three sets of journals: the top five economics journals, the top three political science journals, and

top field journals in economic history.[8] Top-five economics journals include *The American Economic Review*, *The Journal of Political Economy*, *The Quarterly Journal of Economics*, *The Review of Economic Studies*, and *Econometrica*. Top-three political science journals include *The American Political Science Review*, *The American Journal of Political Science*, and *The Journal of Politics*. Our categorization of economic history journals includes *The Journal of Economic History*, *Explorations in Economic History*, *Cliometrica*, *The European Review of Economic History*, and *The Economic History Review*.

For economic history journals, Figure 1 includes all papers that mentioned "natural experiment," "policy experiment," or "quasi-experiment." For general interest generals in economics and political science, we impose the additional condition that the paper must mention "history" or "historical." We then manually exclude papers that appeared to be about contemporary rather than historical natural experiments. The extent to which papers that use the language of natural experiments to study a historical episode are targeted at top journals versus economic history field journals is striking. This disparity is consistent with broader trends in economic history noted in other studies: Abramitzky (2015) notes an increasing share of economic history papers being published in top journals over 1970 to 2010, while Margo (2018) finds that recent cohorts of economic historians are more likely to publish in top journals.

One inference that could be drawn from Figure 1 is that researchers trying to place history papers in top journals face incentives to characterize their studies as natural experiments to increase their odds of success. It appears that this incentive is much less pronounced for cliometrics field journals. In the remainder of this section, we grapple with the implications of the increasing use of "natural experiments" by cliometricians. Section 3.1 discusses the goal of natural experiments in history. Section 3.2 discusses how a deep understanding of the "historical and institutional context" emphasized by Lyons et al. (2007) is an essential complement to statistical techniques for evaluating threats to a valid natural experiment. Section 3.3 discusses why there is a tension facing historical studies aimed at top journals: getting the historical context right is simultaneously essential for properly validating natural experiments while also potentially inconvenient for presenting a "clean" research design to a general audience in economics or political science.

[8] Our count omits economic history papers that utilize natural experiments but do not use these words. For instance, Dippel (2014) uses an IV strategy but does not use the term "quasi-experimental" and is therefore not included in our count despite matching our criterion very well in a qualitative sense.

3.1 The Purpose of Natural Experiments

The goal of any natural experiment is to mimic the features of controlled laboratory experiments conducted by physicists and other scientists (Angrist and Pischke, 2010; Sekhon and Titiunik, 2012). Although natural experiments come in different shapes and sizes that call for different "identification strategies," Diamond and Robinson (2010) outline several key concerns that any researcher should consider when studying a natural experiment from history. The four key issues that social scientists must address in a natural experiment include *(i)* the direction of causality, *(ii)* the influence of omitted variables, *(iii)* non-random assignment of "treatment," and *(iv)* causal mechanisms. The first three issues are geared toward mimicking conditions in the lab, whereas the fourth is somewhat unique to the social sciences.

In the lab, the direction of causation is typically not a concern because the experiment is explicitly designed with a specific causal pathway in mind: a researcher wishes to study the effect of X on Y, and so exposes Y to some change in X. The potential for reverse causality arises in the study of (observational) historical data because it may not be known *ex ante* whether a change in X affected Y or vice versa. Hence, one feature of a valid natural experiment is that it is an environment in which variation in the "treatment" of interest was credibly not caused by underlying variation in the outcome of interest. Often, reverse causality claims can be evaluated temporally (e.g., by demonstrating that the change in X pre-dates the change in Y), particularly in the context of cliometric studies.

Another concern for experiments of any kind is the influence of omitted variables. In some sense, laboratories are designed specifically for the purpose of minimizing the influence of omitted, or confounding, variables by giving the researcher precise control over the conditions of the experiment. Because the real world is not so neat, much of the intuition underlying natural experiments is to find settings where the potential influence of omitted variables is minimized. This logic is evident in each of the common research designs we discuss in Section 3.2. Although a variety of tests exist to help evaluate the potential for omitted variables to compromise a natural experiment, these techniques ultimately rely on a well-established knowledge of the event in question so that the researcher is aware of the full set of potential confounding variables (Dunning, 2008; Angrist and Pischke, 2010; Sekhon and Titiunik, 2012).

Laboratory experiments also help ensure that differences in outcomes can be attributed to differences in treatment by randomly assigning treatment status. Non-random assignment, or "selection into treatment," raises the possibility that differences in observed outcomes may be attributable to underlying, pre-existing differences in the units that were exposed to the treatment vs. those that remained in the "control" group. In studies of observational data, true random assignment is rarely an option. Researchers must attempt to rule out or minimize the potential for non-random assignment of whatever natural experiment generated variation in "treatment" across observational units. Even when assignment to treatment is as good as random, careful attention must be paid to ensure that comparisons across different subpopulations actually identify the effect of interest (Sekhon and Titiunik, 2012).

The fourth and final ingredient for a natural experiment is the exploration of causal mechanisms. Whereas lab experiments can be repeated to confirm the robustness of a relationship, the historical events that give rise to natural experiments are unique (Nunn, 2020). Although a specific event may be hard to replicate, the underlying economic and political forces at work may be quite general. Hence, providing evidence about specific mechanisms connects natural experiments to theory, which has several important benefits. First, checking results against predictions from theory helps to weed out spurious empirical relationships, or "false positives," that do not correspond to casual economic and political forces (Voth, 2020). Second, linking mechanisms to theory bolsters the credibility of a natural experiment by positing general relationships that can be tested and verified in other contexts. Finally, shedding light on mechanisms that likely exist in other contexts also broadens the impact of a given study.

In this paper, we are particularly focused on natural experiments that are based on specific historical policies and institutions (as opposed to, say, geographic or climatic variation). This is because a lack of context-specific historical knowledge about the political economy that generates institutional variation creates more serious threats to validity than does a lack of knowledge in studies reliant on less endogenous geographic or climatic phenomena. As Dunning (2008) puts it: "many of the interventions that might provide the basis for plausible natural experiments in political science are the product of the interactions of actors in the social and political world, and it can strain credulity to think that these interventions are undertaken in ways that are independent of the characteristics of the actors involved, or in ways that do not encourage the actors to 'self-select'

into treatment and control groups in ways that are correlated with the outcome in question." In Dunning's words, discovering potential threads to validity along these lines requires "detailed case-based knowledge often associated with qualitative research."

As we discuss in Section 3.2, this problem is particularly pronounced in cliometrics due to the idiosyncratic nature of many historical studies and the corresponding asymmetry of contextual knowledge between authors and referees. As we discuss in Section 3.3, the costly nature of this style of research creates a temptation to focus solely on statistical diagnostics while overlooking factors that can complicate the empirical design or make it less persuasive.

3.2 The "Historian's Skill" in Validating Research Designs

There are three main classes of natural experiments utilized in cliometrics and in applied economics more broadly: instrumental variables (IV), regression discontinuity (RD), and panel fixed effects (FE) estimators (Angrist and Pischke, 2010).[9] Each of these empirical designs has its own set of statistical diagnostics that can be used to evaluate the underlying identification assumptions. Ultimately, these tool are necessary, but not sufficient to validate a natural experiment. As Angrist and Pischke (2010) put it: "The best of today's design-based studies make a strong institutional case, backed up with empirical evidence, for the variation thought to generate a useful natural experiment." We briefly review each research design's ability to address Section 3.1's four key issues in evaluating natural experiments. We also highlight how each approach is usually applied in historical studies, raising some unique considerations not typically present in studies of modern data. Our concern is that an increasing emphasis on quantitative validation of causal identification strategies has led to an offsetting decrease in the emphasis placed on building a strong institutional case. To illustrate why both are necessary, this section provides several examples of how standard diagnostics can fail to detect important threats to identification when not informed by sufficient historical context.

[9] Angrist and Pischke (2010) list differences-in-differences (DiD) as the third type of empirical design, but we also consider other designs that use panel variation to absorb unobserved differences in cross-sectional units over time even if they are not strictly DiD estimators. Panel fixed effects strategies are sometimes called *generalized* DiD strategies, but increasingly the term *two-way FE* strategy is becoming more common (de Chaisemartin and D'Haultfœuille, 2020).

Instrumental Variables

The core intuition of the IV approach is to find a predictor of treatment status that is otherwise uncorrelated with the outcome of interest and use this "instrument" to generate exogenous variation in the treatment. In practical application, the econometric condition of *exogeneity* can usefully be broken down into two conditions: an instrument needs to be *external* in the sense that it is not subject to selection, not correlated with omitted variables, and not subject to reverse causality. And it needs to be *excludable* in the sense that it affects the outcome being studied only through the endogenous variable of interest (Heckman, 2000; Deaton, 2010). Neither assumption can be tested directly,[10] but researchers can provide both quantitative and qualitative evidence to support their plausibility. Detailed historical knowledge is crucial for validating instruments in cliometrics. There are always a litany of potential exclusion restriction violations, even in studies of modern data. Compared to a study of contemporary policy, the set of unknown factors that could violate the exclusion restriction tends to be larger for studies of historical events because the passage of time creates greater distance between the researcher and the object of study, increasing the possibility for confounding events, policies, and correlations to go unnoticed (Voth, 2020).

One area of cliometrics where the context-specific knowledge of historians may not actually convey a particularly strong comparative advantage in gauging the validity of an IV is the subarea of *persistence studies* that rely on geographic sources of variation. Some of the first economic history papers to grapple with modern experimental design fall in this category: Hall and Jones (1999) use distance to the equator as an instrument for Western European influence when studying the effect of "social infrastructure" on capital accumulation, and Acemoglu et al. (2001) use settler mortality as an exogenous source of variation in colonial institutions to study the effect of institutions on economic growth (both studies compare outcomes across countries). Other historical geographic characteristics (implicit in the case of settler mortality) used as an instruments for a (historical or contemporary) treatment variable include, inter alia, sailing distances (Nunn, 2008), climatic and soil conditions (Alesina, Giuliano, and Nunn, 2013), and mineral deposits (Dippel, 2014). Historians' knowledge may not be particularly essential to gauging if a geographic feature

[10] So-called "over-identification tests" can be conducted when the researcher has more than one instrument for each endogenous variable, however. A failure to reject the null of over-identification provides evidence in support of the exclusion restriction, but the nature of the test precludes a definitive rejection of a potential unknown violation of the exclusion restriction.

satisfies the *excludability* condition as an instrument.[11]

By contrast, an example of a cliometric IV literature where historians' knowledge may be absolutely critical is the large number of studies that use a distance-based instrument. Many cliometric studies use distance to geographic features, infrastructure, or important cities as instruments to identify various institutional changes or decisions in the past. Typically, the exclusion restriction is of the form: "distance to z was critical in the past and so induced important variation in x, but due to subsequent changes in conditions (technology, market forces, political regimes, etc.), z is no longer important. Hence, z should only affect outcome of interest y through its effect on x." While such logic can be appealing, it may be very difficult for a reader unfamiliar with the historical setting to gauge what else may correlate with "distance to z." For instance, consider three important papers by Acemoglu, Johnson, and Robinson (2005), Becker and Woessmann (2009), Dittmar (2011), which respectively use coastal location, distance to Wittenberg, and distance to Mainz to study economic development in European cities between 1500–1900. Anyone with a familiarity of the European map would suspect that these two distance measures are highly (positively) correlated with one another, as well as highly (negatively) correlated with being on a coast. While the correlations between these three measures could of course be estimated quantitatively, the broader point here is that it can be generally difficult for a non-historian to recognize what other important spatial characteristics an instrument such as "distance to z" may correlate with. Once again, the "historian's skill" is likely needed to appreciate the potential for confounding variables or institutions that develop over time to be spatially correlated with the instruments. Lastly, and most pertinently, any IV strategy that is explicitly based on a historical policy or institution will be most in need of scrutiny from a historian who has a deep understanding of said policy or institution.

Regression Discontinuity

Regressions discontinuity (RD) designs exploit sharp differences in exposure to treatment across some threshold to estimate the effect of a policy or institution. RD designs first gained traction in labor economics, where cutoffs in test scores that were subsequently used as eligibility criteria

[11] There are some additional concerns that are specific to IV in persistence studies. For instance, when a historical instrument predicts a *contemporary* treatment, additional interpretational and econometric issues arise around the *excludability* assumption (Casey and Klemp, 2021). When a historical instrument predicts a *historical* treatment, on the other hand, this may considerably amplify the difference between the estimated *Local Average Treatment Effect* (LATE), and the actual object of interest, namely the *Average Treatment Effect* (ATE) (Bisin and Moro, 2020).

generated sharp differences in educational outcomes (such as access to scholarships) for students whose scores were actually quite close (Angrist and Pischke, 2010). The intuition of RD designs is that we can compare units that are extremely similar along some key "running" variable by comparing only those within a narrow bandwidth on either side of an arbitrary cutoff, thus minimizing omitted variable bias (e.g., comparing students who barely passed a test to those that barely failed). Validating such designs typically involves testing for differences in observed covariates on either side of the boundary and ensuring that individuals did not have precise control over the "running" variable and hence could not intentionally sort across the boundary.

In cliometrics, spatial RD designs have become particularly popular. As the name implies, spatial RD designs use space as the "running" variable and compare outcomes of interest just across some (hopefully) arbitrary spatial boundary that typically denotes a sharp change in policy or governing institutions. For example, Libecap and Lueck (2011) exploit the boundary of the historical "Virginia Military District" in Ohio to compare land values of nearby parcels subject to different demarcation systems on either side of the border. While most of Ohio was surveyed under the rectangular Public Land Survey System grid, the Virginia Military District used the older, less uniform "metes and bounds" system of demarcation, resulting in higher transaction costs and lower land values just inside the border. Other examples of spatial RD designs in cliometrics include Dell (2010), who uses the boundaries of the historical *mita* system of labor obligations, Becker, Boeckh, Hainz, and Woessmann (2016), who use the border of the Habsburg Empire that cuts across several present-day Eastern European countries, and Ambrus, Field, and Gonzalez (2020), who use the boundary of the cholera-infected water-pump that caused the 1854 London Cholera outbreak. Spatial RD designs raise some unique validity concerns. Whereas the thresholds in RD designs based on test scores are arbitrarily set *ex ante*, spatial boundaries may themselves be endogenous or correlated with omitted variables. Hence, the validity of a spatial RD design hinges on the researcher's ability to establish that *only* the institution of interest varies at the boundary, that individuals cannot sort themselves across the boundary, and that the boundary itself is exogenous (Keele and Titiunik, 2015).

A variety of statistical diagnostics exist to help validate RD designs by detecting "bunching" of observations near the boundary and testing for differences in control variables on either side (see Lee and Lemieux (2010) for a survey of RD designs and methods). Broadly, the intuition of

all of these tools is to ensure that the population being studied varies smoothly in space so that differences in the outcome of interest can be attributed solely to different institutions. Even with these tools in hand, "considerable substantive knowledge is needed to credibly exploit geographic boundaries as RD designs" (Keele and Titiunik, 2015). This is especially true when the design examines the long-run effects of some historic border discontinuity because there is a longer duration over which threats to validity may arise, and hence a larger set of potentially unobserved confounding institutions (McCauley and Posner, 2015).

One important limitation of quantitative methods for validating spatial RD designs is that they are only as good as the available data measures. As Keele and Titiunik put it: "whether falsification tests based on covariates lend credibility to the design will depend on whether the covariates selected are closely connected to the outcome and the treatment of interest, an issue that will depend on substantive knowledge behind each particular application." This creates a potential problem in the peer review process if the referees are not familiar with the institutions and historical period being studied. Whereas a historically knowledgable referee may anticipate potential confounding variation and push the researcher to collect important covariates, an unfamiliar referee may simply take as given the set of variables presented by the author. Often, the data used in cliometric studies are collected first-hand by the researcher, which means that the validation tests for differences in covariates across the boundary are all conditional on decisions made by the researcher about which variables to construct.

The boundary itself also presents unique challenges for spatial RD designs in cliometrics. In traditional RD designs, the cutoff that generates a policy discontinuity is often truly arbitrary with respect to the running variables (e.g., a score of 70% on an exam) because it is assigned based on a hypothetical distribution of outcomes before those outcomes are observed. With a spatial RD, the running variable (space) is always known and observable prior to the formation of the boundary, raising the possibility that the boundary itself is endogenous with respect to any variable that varies over space. This concern is especially problematic in cliometrics because the endogenous drivers of the historical boundary may no longer be observable to the researcher. If those drivers of the boundary also affect the outcome of interest, identification is compromised. Standard RD diagnostics are unlikely to adequately address endogenous boundaries unless the researcher engages deeply with the history of the policy and context in question to ascertain po-

tential determinants of the boundary. For example, a number of political science studies attempt to exploit the relatively arbitrary nature of national boundaries in Africa in RD designs to study the effect of historical policies on ethnic groups that span these boundaries (Firmin-Sellers, 2000; Miles and Rochefort, 1991; Coast, 2002; Miles, 2005). However, McCauley and Posner (2015) argue that that these designs are often invalidated by various omitted variables from history including pre-partition settlement patterns, previous military conquests, and boundary negotiations that explicitly took account of demographic factors.

Another potential problem for spatial RD designs in cliometrics arises from the interpretation of results. Without deep contextual knowledge about whether/how a policy was enforced and for what period a time, the researcher runs the risk of presenting statistically sound results and interpreting them incorrectly.[12] For example, Arroyo Abad and Maurer (2019) illustrate that the long-run effects of Peru's mining *mita* documented by Dell (2010) are fairly specific to the Huancavelica and Potosi *mitas* used by Dell to construct an RD, but do not necessarily generalize to the broader set of communities subject to the *mita*. In part, this stems from differences in the historical sources used to determine the treatment status of specific areas. More broadly, Arroyo Abad and Maurer (2019) argue that, unlike for Huancavelica and Potosi, "the *mita* was negotiable for communities that were sufficiently wealthy, strategic, or stubborn in their resistance" and that many individuals were able to avoid exposure to the *mita* by migrating (i.e., sorting across the boundary).

Even when results are not compromised by endogenous boundaries or sorting, they may be falsely attributed to the wrong causal mechanism if the researcher lacks a sufficient depth of knowledge regarding the actual implementation of the historical policy under study or if there are multiple "compound treatments" that occur along the discontinuity that are not appreciated by the author (Keele and Titiunik, 2015). Diamond and Robinson's emphasis on documenting mechanisms can help alleviate this problem by prompting the researcher to search out additional evidence that is consistent with their interpretation of the results. Our broader point is that the historian's skill is especially crucial for constructing valid RD designs in history precisely because these designs can pass a variety of quantitative diagnostics and appear quite sharp even if the

[12] With spatial comparisons, just as with time-series analysis, there is also always the possibility of spurious correlation (Kelly, 2019).

underlying design is invalid or the findings are falsely attributed to the wrong institutions.

Panel Fixed Effects

Finally, panel data methods are widely used empirical designs that utilize repeated observations of individuals, states, or countries to study how outcomes of interest change over time in response to changes in some treatment of interest. In general, a key advantage of panel data is the ability to include fixed effects for each cross-sectional unit that purge any time-constant differences between observations (similarly, time fixed effects can be used to purge common shocks that affect all units simultaneously). This approach helps address omitted variable bias from studies of the effect of a time-varying institution on some time-varying outcome. To address potential selection and reverse causality in these time-varying relationships, difference-in-difference (DiD) designs compare differences in a treated vs. control group before vs. after some treatment or key institutional change. Examples of DiD in the broader applied econometric literature are ubiquitous (see Angrist and Pischke (2010) for some headline examples).

The use of DiD and fixed effect (FE) estimators in cliometrics is perhaps less widespread than IV and RD because assembling repeated historical observations can be challenging. Often, researchers are lucky to construct a single cross-section of observations from a historical period and then relate variation in this cross-section to modern cross-sectional variation. Regularly published historical censuses are the primary example of historical panel data, allowing researchers to study county-level or state-level data over time. More recent advances in record linking have also allowed researchers to construct individual-level panel data by linking records of the same individual across multiple census waves.[13]

Many FE and DiD designs in cliometrics focus on United States history, and on the 20th century in particular. For example, Goldin and Katz (2002) exploit state-level changes in teenagers' access to contraceptives driven by arguably exogenous changes in the "age of majority" induced by concerns related to the Vietnam War military draft. By comparing differences across states that did or did not change the age of majority (the first difference) before vs. after the change took effect

[13] Another area where there have been major advances in historical panel data is dealing with changing boundaries/definitions of cross-sectional units over time, which create challenges in comparing data across time periods. The changing boundaries of U.S. counties over time, for instance, used to create major headaches for researchers but are now routinely dealt with.

(the second difference), Goldin and Katz find that access to contraception reduced the probability of marriage before the age of 23 for women born between 1935 and 1957 who were teenagers when the policy changes took effect.

As with spatial RD designs, there are a growing number of quantitative exercises that can be used to validate panel methods and DiD designs in particular. A common approach is to test for "pre-trends," or differences between groups exposed to different policies or institutions *before* differences in exposure occur. There has also been an explosion of work in recent years that has produced a variety of techniques for diagnosing, characterizing, and addressing treatment effect heterogeneity across different groups that receive treatments at different times (Goodman-Bacon, 2018; de Chaisemartin and D'Haultfœuille, 2020; Steigerwald, Vazquez-Bare, and Maier, 2020).

Another important concern with DiD designs is the validity of the "stable unit treatment values assumption" (SUTVA). SUTVA is violated if there are spillovers between treated and untreated cross-sectional units; in other words if the "treatment" of interest affects not only the treated but also the untreated observational units. SUTVA is important because the basic logic of DiD requires stability in the control group in order to make sense of relative changes observed in the treatment group. Unlike the concern about pre-trends, SUTVA typically cannot be verified quantitatively. This concern applies to all DiD designs, and is not specific to cliometrics. Moreover, Lyons et al.'s pillar *(ii)* — the use of theory — is arguably the most useful skill for addressing potential SUTVA violations. Some of the structural approaches mentioned in Section 2 do exactly this. For example, Donaldson and Hornbeck (2016) use a structural approach to explicitly model spillovers between counties when studying the effect of railroads on economic development, arguing that this is an improvement on previous DiD studies that were subject to potential SUTVA violations due to these spillovers (e.g., Atack, Bateman, Haines, and Margo (2010), Atack and Margo (2011), and others).

A third important concern with DiD designs is dynamic omitted variable bias (OVB), which arises if, for instance, a measured policy change is accompanied by other unmeasured changes that differentially impact the observational units that are treated by the policy. This concern is again not specific to cliometrics. However, as we argued before, the threat that OVB poses to the validity of research designs is more pronounced in cliometrics than in modern-day settings precisely because the context-specific knowledge needed to address it is more likely to be under-

supplied in historical research.

Another factor that amplifies OVB concerns is that historical panel data are often observed at very low frequency. One example of this is the large number of studies drawing on the decennial U.S. Census. Another even more stark example is a large number of studies that use Bairoch's (1991) population data for European cities from 1000–1900 (in lieu of nonexistent GDP and income measures), see e.g. Acemoglu et al. (2005), Dittmar (2011), Nunn and Qian (2011). These studies use outcome data that are only measured every 50 to 100 years. This low frequency of long-run panel analyses increases the possibility of various unobserved trends and OVB between observed data points — identification in these settings requires that there are no unmeasured changes that differentially impact the treatment and control groups over the course of decades, if not centuries. This makes the historian's context-specific knowledge absolutely critical for validating panel designs with "long differences."

3.3 The Tension

Ideally, all cliometric studies would take seriously Lyons et al.'s fourth pillar of "the use of a historian's skill to both judge source material and place it in its institutional and historical context" when validating research designs. However, there is a tension associated with the potential use of contextual knowledge to validate empirical designs. Sufficient familiarity with historical details can bolster a researcher's case for validity, but it can also potentially undermine an otherwise "clean" natural experiment by revealing inconvenient facts about the way a policy was actually implemented. A so-called natural experiment can appear less exogenous as one learns more about the historical details surrounding the variation being studied, including the political economy of the policy, potential gaps in enforcement, other related policies, etc.

The benefits of characterizing institutional variation in overly clean and simplistic terms are not unique to cliometrics — the credibility revolution has created a premium for elegant empirical designs that exploit plausibly exogenous variation, especially at top journals. What *is* somewhat unique to cliometrics is the idiosyncratic nature of many historical studies and the corresponding asymmetry of contextual knowledge between authors and referees. Many fields in economics are characterized by a large set of researchers studying a relatively small set of core questions or policies that the entire field has familiarity with because the field is defined around a unified topic.

Drawing on the example given in the introduction, regulation of pollution is a ubiquitous topic in environmental economics, and nearly every graduate student in this field becomes familiar with a litany of studies on the Clean Air Act in their field courses. Cliometrics is just the opposite: because the only common denominator is history, there are a relatively small number of researchers studying any given historical period or episode (and even fewer who are not co-authors). The upshot is that a referee's ability to evaluate a paper in cliometrics is often constrained, at least to some degree, by the contextual details provided by the researcher.

Researchers' incentives to downplay inconvenient historical details are partially kept in check by the norms of the field itself. From the Nobel Prize committee to Lyons et al.'s survey, cliometricians have always emphasized the importance of taking history seriously, and more recent work has emphasized the importance of contextual knowledge for validating historical studies (Nunn, 2020). The biggest incentive problem exists for cliometrics studies targeted at top general-interest generals. Among journals such as the "top five" in economics, the bar for causal identification is higher and the likelihood of a referee with contextual knowledge is simultaneously lower, thus encouraging researchers to prioritize a narrative with a clean research design. Abramitzky (2015)'s retelling of Gavin Wright's anecdote that we mention in the introduciton is a case is point.

We do not mean to suggest that studies that get the history wrong do so intentionally. Instead, we are calling attention to the fact that market forces within cliometrics — the demand for sharp identification and the limited supply of historically-informed referees on a given topic — create bad incentives as well as steep tradeoffs between time spent on econometric vs. historical methodology, especially for research aimed at top general-interest journals. And of course, the level of engagement with history is not a binary — the tendency to gloss over institutional details certainly exists on a spectrum. Although it is possible that some researchers learn inconvenient facts about the institutions they are studying and subsequently sweep these facts under the rug, we hope this is not the case.

We believe the more pervasive problem is that researchers may not deeply engage with the history in the first place. One can learn the basic contours of a policy (the "where and when" of implementation) and move on to constructing a statistically compelling empirical design without ever really learning about the political economy (the "how and why" of implementation) that might invalidate the research design or dramatically change its interpretation. In this way, would-

be cliometricians remain blissfully ignorant of inconvenient facts, as do their referees. Again, such a strategy is more likely to result in a swift rejection in other fields where the pool of common knowledge is more concentrated (this strategy would not work for studying the Clean Air Act, for instance). This practice is not dishonest *per se*, and it is consistent with the incentives generated by the current institutional environment, but it does not make for especially good cliometrics. This is perhaps a symptom of the gradual dissolution of cliometrics as a methodologically distinct field, as discussed by Margo (2018).

4 Keeping the "Clio" in Cliometrics

In Section 3, we contrast the enduring importance of historical contextual knowledge in cliometrics with systematic incentives that potentially generate a trend toward studies of historical natural experiments that only engage with history superficially. Ultimately, these problems stem from a shift in emphasis away from historical knowledge and toward statistical techniques. We think that there is no shortage of qualified critics of econometric methodology in economics and political science. The question, then, is how to ensure we retain the "a historian's skill" in the field of cliometrics.

How might cliometricians be encouraged and incentivized to utilize a historian's skill when studying natural experiments? Clearly, the possibility always exists to seek out critical feedback from historians expertise on a particular topic. The problem, we believe, is that the current set of incentives facing cliometricians does not reward such an exercise. We propose correcting the incentive problem discussed in Section 3.3 through changes in the refereeing process of general-interest journals handling econometric papers. Researchers' incentives for what to include in a paper are dictated by the editorial review process. The ever-increasing length of robustness checks and appendix exercises are a symptom of the increasingly stringent empirical standards applied at many journals. Hence, the most direct way to incentivize cliometricians to engage more thoroughly with history is to alter or amend the review process itself, at least as it pertains to historical (or other context-heavy) works in general-interest journals. As we discuss in Section 3.3, referees of cliometric works in general-interest journals are more likely to be selected for their technical expertise, and perhaps for working on *economically* related topics, than for their historical and

contextual knowledge. As such, it may be difficult for them to evaluate whether a given paper is historically accurate or glossing over important details.

The set of referees who possess both technical expertise and the relevant context-knowledge may be very small or even empty. In such cases, we believe editors should have a lexicographic preference for ensuring at least a modicum of scrutiny grounded in contextual knowledge is involved in the refereeing process. If experts with the relevant historical knowledge are too time-constrained, or not technically trained to provide a full report commenting on the parts of the paper devoted to statistical identification, then we believe that editors should instead request a shortened report that only comments on the veracity and plausibility of the historical natural experiment as it is described in the paper.

To facilitate such an approach we propose that authors submitting a cliometric paper to a general-interest (or perhaps any) journal include a separate "historical abstract" that delineates the purely historical parts of the paper. The historical abstract would consist of the researcher's summary of the policy, institution, or other variation that they claim comprises a natural experiment. This summary would include a characterization of the source of variation itself, as well as plain-English statements of the identifying assumptions necessary for validity of the research design. Historical details could be in part provided in an in-depth historical appendix, since it is probably both infeasible and undesirable to insist that every paper include a lengthy historical background section.[14] Multiple appendices with empirical robustness checks are now commonplace in cliometrics, and the addition of a supplemental historical appendix would allow researchers to demonstrate that they hold a sufficient understanding of the setting under study without unduly lengthening main text of manuscripts.

Soliciting targeted feedback from historians on "historical abstracts" would also solve two potential supply-side concerns associated with our proposal. The first concern is that historians may simply be unwilling to participate in the referee process for economics journals. Limiting the scope of feedback to a short abstract reduces the costs to potential referees, but our proposal still requires a degree of altruism from those referees (as does all peer review). The second concern is that many historians may find even relatively careful cliometric studies to be unsatisfactory

[14] One of the editors of this journal, Scott Gelbach, recently advocated for this approach in a blog post about facilitating more cooperative interaction between historians and social scientists. See: https://broadstreet.blog/2020/09/23/navigating-the-frontier-between-history-and-social-science/.

based on methodological differences between history vs. more quantitative social science that have only grown over time (Diamond and Robinson, 2010; Margo, 2018). Historical abstracts help address the latter concern by narrowing the scope of criticism to the historical accuracy of specific statements about past policies and events that correspond to the researcher's identifying assumptions. For example, a historian can evaluate the veracity of a claim like "people were not able to freely migrate across boundary X" even if they doubt the benefit of conducting an RD design in the first place.

Our proposal is similar in spirit to the recent addition of "data editors" that are responsible solely for verifying the replicability of studies at many top economics journals.[15] Just as researchers are now required to submit their data and code for replication, editors reviewing historical papers could require authors to submit a historical abstract along with suggestions for subject matter experts on the historical period in question. We believe our proposal would ensure that the trend toward ever-increasing scrutiny of quantitative empirical work does not come at the expense of the "historian's skill." Our proposal imposes some additional costs on researchers, but would ultimately also benefit them, as it offers a clearer path for bolstering empirical designs by providing additional historical evidence to complement quantitative evidence. In settings where the quantitative evidence for validity is inconclusive, historical information might provide a powerful supplement.

5 Conclusion

In this article, we review the chronological evolution of cliometrics from its early emphasis on quantification and theory to the more modern focus on causal inference. We discuss complementarities and potential tradeoffs with the contextual historical knowledge that was a hallmark of the work of Fogel, North and others. As a conceptual framework, we use Lyons et al.'s characterization of four distinctive pillars of cliometrics: *(i)* the use of quantifiable evidence, *(ii)* the use of theoretical concepts and models, *(iii)* the use of statistical inference, and *(iv)* the use of a historian's skill to both judge source material and place it in its institutional and historical context. We discuss how the post-1990s credibility revolution has led to the increased prominence of natural experi-

[15] This analogy is an apt one, as Nunn (2020) suggests that one solution to the lack of replicability in historical studies is for authors to closely cross-reference their findings with the historical literature.

ments in cliometrics, and how this has come at a cost to the application of the "historian's skill" in cliometric studies. We discuss the reasons for this, with particular emphasis on the incentives created in the refereeing process in general-interest journals. We propose an easy-to-implement amendment to best practices in the refereeing process of historical (or other context-heavy) works in general-interest journals.

The proposal we outline imposes costs on researchers, editors, and referees, so it is natural to ask whether the potential benefits are worth the trouble. Narrowly speaking, the benefit of keeping the "clio" in cliometrics is to safeguard the validity of quantitative studies of history, which have increasingly emphasized causal inference. The benefits of an improved understanding of causal relationships in social science generally and in history in particular have been well-articulated by others (Angrist and Pischke, 2010; Diamond and Robinson, 2010). Our contribution, we hope, is to demonstrate the importance of contextual knowledge for ensuring that these benefits are real rather than illusory. False confidence in the results of invalid natural experiments may do more harm to scientific knowledge than would a measured interpretation of quantitate findings that make no claims to causality.

Creating incentives and structures for more interaction among cliometricians and historians could facilitate new interdisciplinary collaborations, yielding benefits for both groups. In addition to the necessary check on validity emphasized throughout this paper, historians' contextual knowledge promises to be a virtual treasure trove of potential empirical designs for cliometricians to explore — for every "not-so-natural" experiment generated by a superficial reading of history, we suspect there are many more that go unnoticed by the same lack of attention to historical detail. On the other hand, historians can benefit from cliometricians' skill in bringing new forms of evidence to bear on old questions, as emphasized in Section 2. Increased dialogue between historians and cliometricians offers to generate new research designs, new data sets, and new puzzles to be solved that would otherwise be missed by one group in isolation from the other.

Finally, misplaced confidence in the causal interpretation of cliometric studies can also have implications outside the academy. Cliometricians have a long tradition of turning to history as a laboratory for understanding different types of policy, and studies of historical policies and their persistence are increasingly used to inform contemporary policy (Nunn, 2020). Hence, "false positives" and poorly understood mechanisms may lead to bad policy in addition to bad history.

Even when they are internally valid, studies that analyze policies divorced from their historical context threaten to give the false impression that the same policy can be transplanted into a new context with similar results. The policy reforms associated with the "Washington Consensus" for promoting economic development globally led to disappointing results for exactly this reason (Alston, 2017). When it comes to repeating past mistakes, learning the wrong lessons from history may be just as bad as not learning at all.

References

Abramitzky, R. (2015). Economics and the Modern Economic Historian. *The Journal of Economic History 75*(4).

Acemoglu, D., S. Johnson, and J. Robinson (2005). The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth. *American Economic Review 95*(3), 546–579.

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review 91*(5), 1369–1401.

Alesina, A., P. Giuliano, and N. Nunn (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics 128*(2), 469–530.

Allen, D. W. and P. T. Leeson (2015). Institutionally Constrained Technology Adoption: Resolving the Longbow Puzzle. *The Journal of Law and Economics 58*(3), 683–715.

Alston, L. J. (2017). Beyond Institutions: Beliefs and Leadership. *The Journal of Economic History 77*(2), 353–372.

Alston, L. J. and J. P. Ferrie (1993). Paternalism in Agricultural Labor contracts in the US South: Implications for the growth of the Welfare State. *American Economic Review*, 852–876.

Ambrus, A., E. Field, and R. Gonzalez (2020). Loss in the Time of Cholera: Long-Run Impact of a Disease Epidemic on the Urban Landscape. *American Economic Review 110*(2), 475–525.

Angrist, J. D. and J.-S. Pischke (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Arroyo Abad, L. and N. Maurer (2019). The Long Shadow of History? The Impact of Colonial Labor Institutions on Economic Development in Peru. *The Impact of Colonial Labor Institutions on Economic Development in Peru (August 03, 2019)*.

Atack, J., F. Bateman, M. Haines, and R. A. Margo (2010). Did Railroads Induce or Follow Economic Growth? Urbanization and Population Growth in the American Midwest, 1850-1860. *Social Science History*, 171–197.

Atack, J. and R. A. Margo (2011). The Impact of Access to Rail Transportation on Agricultural Improvement: The American Midwest as a Test Case, 1850–1860. *Journal of Transport and Land Use 4*(2), 5–18.

Auffhammer, M., A. M. Bento, and S. E. Lowe (2009). Measuring the Effects of the Clean Air Act Amendments on Ambient PM10 Concentrations: The Critical importance of a Spatially Disaggregated Analysis. *Journal of Environmental Economics and Management 58*(1), 15–26.

Auffhammer, M., A. M. Bento, and S. E. Lowe (2011). The City-Level Effects of the 1990 Clean Air Act Amendments. *Land Economics 87*(1), 1–18.

Bairoch, P. (1991). *Cities and Economic Development: from the Dawn of History to the Present*. University of Chicago Press.

Bazzi, S., M. Fiszbein, and M. Gebresilasse (2020). Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States. Technical report, National Bureau of Economic Research.

Beard, C. A. (1913). An Economic Interpretation of the Constitution of the United States. Reprint, with New Introduction.

Becker, S. O., K. Boeckh, C. Hainz, and L. Woessmann (2016). The Empire is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy. *The Economic Journal 126*(590), 40–74.

Becker, S. O. and L. Woessmann (2009). Was Weber Wrong? A Human Capital Theory of Protestant Economic History. *The Quarterly Journal of Economics 124*(2), 531–596.

Bento, A., M. Freedman, and C. Lang (2015). Who Benefits from Environmental Regulation? Evidence from the Clean Air Act Amendments. *Review of Economics and Statistics 97*(3), 610–622.

Bisin, A. and A. Moro (2020). LATE for History. Technical report.

Busse, M. R. and N. O. Keohane (2007). Market Effects of Environmental Regulation: Coal, Railroads, and the 1990 Clean Air Act. *The RAND Journal of Economics 38*(4), 1159–1179.

Casey, G. and M. Klemp (2021). Instrumental Variables in the Long Run. *Journal of Development Economics, Conditionally Accepted*.

Chan, G., R. Stavins, R. Stowe, R. Sweeney, et al. (2012). The so2 Allowance-Trading System and the Clean Air Act Amendments of 1990: Reflections on 20 Years of Policy Innovation. *National Tax Journal 65*(2), 419–452.

Chay, K. Y. and M. Greenstone (2005). Does Air Quality Matter? Evidence from the Housing Market. *Journal of Political Economy 113*(2), 376–424.

Coast, E. (2002). Maasai Socioeconomic Conditions: a Cross-Border Comparison. *Human ecology 30*(1), 79–105.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature 48*(2), 424–55.

Dell, M. (2010). The Persistent Effects of Peru's Mining Mita. *Econometrica 78*(6), 1863–1903.

Dennison, T. and S. Ogilvie (2014). Does the European Marriage Pattern Explain Economic Growth? *The Journal of Economic History*, 651–693.

Diamond, J. and J. A. Robinson (2010). *Natural Experiments of History*. Harvard University Press.

Dippel, C. (2014). Forced Coexistence and Economic Development: Evidence from Native American Reservations. *Econometrica 82*(6), 2131–2165.

Dittmar, J. E. (2011). Information Technology and Economic Change: the Impact of the Printing Press. *The Quarterly Journal of Economics 126*(3), 1133–1172.

Donaldson, D. and R. Hornbeck (2016). Railroads and American Economic Growth: A Market Access Approach. *The Quarterly Journal of Economics 131*(2), 799–858.

Dunning, T. (2008). Improving Causal Inference: Strengths and Limitations of Natural Experiments. *Political Research Quarterly 61*(2), 282–293.

Eichengreen, B. (1994). The Contributions of Robert W. Fogel to Economics and Economic History. *The Scandinavian Journal of Economics 96*(2), 167–179.

Firmin-Sellers, K. (2000). Institutions, Context, and Outcomes: Explaining French and British Rule in West Africa. *Comparative Politics*, 253–272.

Fogel, R. W. (1964). *Railroads and American Economic Growth*. Johns Hopkins Press Baltimore.

Fogel, R. W. (1967). The Specification Problem in Economic History. *Journal of Economic History*, 283–308.

Fogel, R. W. and S. L. Engerman (1974). Time on the Cross: The Economics of American Negro Slavery. *Boston: Little, Brown 49*, 170–71.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence from US Daily Newspapers. *Econometrica 78*(1), 35–71.

Goldin, C. and L. F. Katz (2002). The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions. *Journal of Political Economy 110*(4), 730–770.

Goodman-Bacon, A. (2018). Difference-in-Differences with Variation in Treatment Timing. Technical report, National Bureau of Economic Research.

Greenstone, M. (2002). The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures. *Journal of political economy 110*(6), 1175–1219.

Greenstone, M. (2003). Estimating Regulation-Induced Substitution: The Effect of the Clean Air Act on Water and Ground Pollution. *American Economic Review 93*(2), 442–448.

Greif, A. (1993). Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition. *The American Economic Review*, 525–548.

Greif, A., P. Milgrom, and B. R. Weingast (1994). Coordination, Commitment, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy 102*(4), 745–776.

Hall, R. E. and C. I. Jones (1999). Why do some Countries Produce so much more Output per worker than Others? *The Quarterly Journal of Economics 114*(1), 83–116.

Heckman, J. J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *The Quarterly Journal of Economics 115*(1), 45–97.

Hubbell, B. J., R. V. Crume, D. M. Evarts, and J. M. Cohen (2010). Policy Monitor: Regulation and Progress Under the 1990 Clean Air Act Amendments. *Review of Environmental Economics and Policy 4*(1), 122–138.

Isen, A., M. Rossin-Slater, and W. R. Walker (2017). Every Breath you TakeEvery Dollar youll Make: The Long-Term Consequences of the Clean Air Act of 1970. *Journal of Political Economy 125*(3), 848–902.

Jaworski, T. (2020). Specification and Structure in Economic History. *Explorations in Economic History 77*, 101343.

Jaworski, T. and C. T. Kitchens (2019). National Policy for Regional Development: Historical Evidence from Appalachian Highways. *Review of Economics and Statistics 101*(5), 777–790.

Keele, L. J. and R. Titiunik (2015). Geographic Boundaries as Regression Discontinuities. *Political Analysis 23*(1), 127–155.

Kelly, M. (2019). The Standard Errors of Persistence. *Available at SSRN 3398303*.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny (1997). Legal Determinants of External Finance. *The Journal of Finance 52*(3), 1131–1150.

Lange, I. and A. S. Bellas (2007). The 1990 Clean Air Act and the Implicit Price of Sulfur in Coal. *The BE Journal of Economic Analysis & Policy 7*(1).

Lee, D. S. and T. Lemieux (2010). Regression Discontinuity Designs in Economics. *Journal of economic literature 48*(2), 281–355.

Libecap, G. D. and D. Lueck (2011). The Demarcation of Land and the Role of Coordinating Property Institutions. *Journal of Political Economy 119*(3), 426–467.

Lyons, J. S., L. P. Cain, and S. H. Williamson (2007). *Reflections on the Cliometrics Revolution: Conversations with Economic Historians*, Volume 38. Routledge.

Margo, R. A. (2018). The Integration of Economic History into Economics. *Cliometrica 12*(3), 377–406.

McCauley, J. F. and D. N. Posner (2015). African Borders as Sources of Natural Experiments Promise and Pitfalls. *Political Science Research and Methods 3*(2), 409–418.

McCloskey, D. N. (1985). *The Rhetoric of Economics*. Univ of Wisconsin Press.

McGuire, R. A. and R. L. Ohsfeldt (1989). Self-Interest, Agency Theory, and Political Voting Behavior: The Ratification of the United States Constitution. *The American Economic Review 79*(1), 219–234.

Miles, W. F. (2005). Development, not Division: Local versus External Perceptions of the Niger-Nigeria Boundary. *Journal of Modern African Studies*, 297–320.

Miles, W. F. and D. A. Rochefort (1991). Nationalism versus Ethnic Identity in Sub-Saharan Africa. *The American Political Science Review*, 393–403.

North, D. (1958). Ocean Freight Rates and Economic Development 1750-1913. *The Journal of Economic History 18*(4), 537–555.

North, D. C. (1961). The United States in the International Economy, 1790–1950. *American Economic History*, 181–206.

North, D. C. (1977). The New Economic History After Twenty Years. *American Behavioral Scientist 21*(2), 187–200.

North, D. C. (1981). *Structure and Change in Economic History*. Norton.

North, D. C. (1990). *Institutions, Institutional Change, and Economic Performance*. New York: Cambridge University Press.

North, D. C. (1991). Institutions. *Journal of Economic Perspectives 5*(1), 97–112.

North, D. C. (2005). *Understanding the Process of Institutional Change*. Princeton, NJ: Princeton University Press.

Nunn, N. (2008). The Long-Term Effects of Africa's Slave Trades. *The Quarterly Journal of Economics 123*(1), 139–176.

Nunn, N. (2009). The Importance of History for Economic Development. *Annu. Rev. Econ. 1*(1), 65–92.

Nunn, N. (2020). The Historical Roots of Economic Development. *Science 367*(6485).

Nunn, N. and N. Qian (2011). The Potato's Contribution to Population and Urbanization: Evidence from a Historical experiment. *The Quarterly Journal of Economics 126*(2), 593–650.

Puga, D. and D. Trefler (2014). International Trade and Institutional Change: Medieval Venices Response to Globalization. *The Quarterly Journal of Economics 129*(2), 753–821.

Royal Swedish Academy of Sciences (1993, Oct). The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1993.

Schennach, S. M. (2000). The Economics of Pollution Permit Banking in the Context of Title IV of the 1990 Clean Air Act Amendments. *Journal of Environmental Economics and Management 40*(3), 189–210.

Sekhon, J. S. and R. Titiunik (2012). When Natural Experiments are Neither Natural nor Experiments. *American Political Science Review*, 35–57.

Spirling, A. (2012). US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911. *American Journal of Political Science 56*(1), 84–97.

Steigerwald, D., G. Vazquez-Bare, and J. Maier (2020). Measuring Heterogeneous Effects of Environmental Policies using Panel Data. *Journal of the Association of Environmental and Resource Economists*.

Stone, L. (1979). The Revival of Narrative: Reflections on a New Old History. *Past & Present* (85), 3–24.

Voigtländer, N. and H.-J. Voth (2013). How the West "Invented" Fertility Restriction. *American Economic Review 103*(6), 2227–64.

Voth, H.-J. (2020). Persistence: Myth and Mystery. *Handbook of Historical Economics, Amsterdam: Elsevier North Holland*.

Walker, W. R. (2013). The Transitional Costs of Sectoral Reallocation: Evidence from the Clean Air Act and the Workforce. *The Quarterly Journal of Economics 128*(4), 1787–1835.