Bridging the Qualitative-Quantitative Divide: Best Practices in the Development of Historically Oriented Replication Databases

Evan S. Lieberman

Department of Politics, Princeton University, Princeton, New Jersey 08544;
email: esl@princeton.edu

Annu. Rev. Polit. Sci. 2010. 13:37–59

First published online as a Review in Advance on November 30, 2009

The *Annual Review of Political Science* is online at polisci.annualreviews.org

This article's doi:
10.1146/annurev.polisci.12.041007.155222

1094-2939/10/0615-0037$20.00

Key Words

methodology, validity, measurement, comparative-historical analysis, replication

Abstract

The proliferation of historically oriented replication data has provided great opportunities for political scientists to develop and to test theories relevant to a range of macrohistorical phenomena. But what is the quality of such data? Are the codings or quantitative mappings of historical events, processes, and unit characteristics based on sufficiently solid foundations equivalent to those found in detailed case studies? This article evaluates a set of the most transparently disseminated replication datasets across a variety of research domains from the perspective of best-practice qualitative-historical research. It identifies a wide range of practices, highlighting both fundamental and innovative standards that might be adopted in future research.

INTRODUCTION

Since the publication of a series of seminal works that raised questions about potential pitfalls in case-study research (Achen & Snidal 1989, Geddes 1990, and King et al. 1994), political scientists have been engaged in a largely productive discussion of standards concerning scientific best practice. Although scholars may disagree about which lessons from statistical theory ought to be applied to case-study or "small-*n*" research (see contributions in Brady & Collier 2004), qualitative researchers are surely more self-conscious about their approaches to case selection and causal inference as a result of this exchange.

Now that a great deal of quantitative political science research has turned to macrocomparative historical analysis, particularly through the use of time-series cross-sectional analyses (Beck 2001), there is good reason to investigate how the practice of quantitative-historical research might gain insights from the more qualitative tradition of the discipline. A large body of qualitative methodology research has focused on the practices and pitfalls associated with macrohistorical analysis, which confronts unique challenges owing to the heterogeneity of units and source materials associated with such work. In order to make theoretically valid comparisons, scholars are routinely forced to make contextually sensitive judgments in the scoring of variables (Collier et al. 2004), and such challenges are multiplied in the context of ambitious efforts to classify large numbers of countries or other units for long periods of time. This article revisits a set of methodological concerns that have been raised, if not always solved, by small-*n* researchers, and highlights their relevance to a large and growing body of quantitative research. Specifically, it is concerned with identifying best practices of comparative-historical research, aiming to maximize transparency and measurement validity (Adcock & Collier 2001), which are fundamental principles of good social science research.

In the subfields of comparative politics and international relations, quantitative analyses of time-series cross-sectional data, frequently of the form describing country units over time, have become ubiquitous and, in many cases, highly influential for addressing core disciplinary questions. Such analyses are analogous to what qualitative researchers have long referred to as "macrocomparative" or "comparative-historical" analysis (Skocpol & Somers 1980, Mahoney & Rueschemeyer 2003) in that they seek to generate causal inferences about the determinants of large-scale political phenomena through dynamic comparisons across time and space. The more recent proliferation of historically oriented quantitative analyses can be explained at least in part by broadside charges that qualitative comparative-historical research ought to be more theoretically ambitious in scope (e.g., Kiser & Hechter 1991) and also by the increasing availability of time-varying, country-level replication data. In turn, the proliferation of the latter can be explained by increased ease of sharing via the internet, the advocacy of greater transparency and replication norms (e.g., Herrnson 1995, King 1995), and the development of several large-scale research projects aimed at generating quantitative indexes relevant to a range of phenomena of interest to political scientists.

The central challenge in this review is to critically assess current practices for developing and disseminating the quantitative counterparts to the qualitative tradition of macrocomparative research. Others have already described potential problems of reliability and validity in datasets that are routinely analyzed by political scientists. Herrera & Kapur (2007), for example, identify concerns about validity, coverage, and accuracy, and focus on actor-induced problems in "data supply chains." Several other studies have addressed concerns about the validity of various data-gathering efforts within specific research domains, as will be discussed below. This review complements such analyses by identifying the methodological concerns and best practices identified by case-study researchers that are also relevant to the development of larger-scale datasets. Whereas Mahoney & Goertz (2006, pp. 244–45)

describe two "culturally" distinct approaches to conceptualization and measurement across qualitative and quantitative research traditions, my goal here is to identify ways in which some of the more "virtuous" traditions have been or could be transferred.

I begin by introducing a set of key methodological concerns that have been raised about historically oriented research more generally, particularly among "case-oriented" or small-*n* scholars. As distinct from historians, political scientists have a primary interest in analytically equivalent comparisons, especially the observation of negative and contrasting outcomes. Whether such work is identified as "analytic narratives" (Bates et al. 1998) or comparative-historical analysis (Mahoney & Rueschemeyer 2003), political scientists have dug into the historical record and provided narrative accounts in order to develop and test broader theoretical propositions of political relationships. Historically oriented scholars have routinely confronted a set of methodological challenges in such investigations.

Second, I consider how such methodological concerns also pertain to a set of historically oriented quantitative data-collection efforts. These projects have involved the translation of narrative records of events and processes into numerical scores in the form of indexes, scales, or dummy variables indicating the presence or absence of some trait, such as regime type or degree of conflict. (I am not concerned here with the generation of count data such as numbers of battlefield casualties or dollars per capita.) I consider these concerns only in the context of those data-collection projects that have made a substantial effort to disseminate some of the underlying facts or narratives associated with the quantitative mappings. In this sense, this review is not representative of the field; it is severely biased toward the best and most transparent replication data available. I label these endeavors historically oriented and integrated replication databases (HIRDs). They are systematically collected and theoretically informed containers of facts and observations for a consistent set of units over time. By disentangling narrative and quantitative historical data from the principal research outputs for which they were designed, HIRDs offer political scientists unique opportunities to assess the quality of completed research, to add new information, and to generate independent analyses based on prevetted data. They promise to promote the accumulation of social scientific knowledge with a level of transparency that allows for updating and corrections in light of new research findings.

HIRD: historically oriented and integrated replication database

The HIRDs discussed in this article report on macro-level units or public officials, subjects about which scholars may obtain additional information to contribute to and refine the original database. Along these lines, I consider here HIRDs that are publicly available via the internet. I do not include "encyclopedic" databases, such as the CIA World Factbook or Wikipedia. Although these may contain a wide range of quantitative and qualitative information about units relevant to political scientists, and the information is often presented in a consistent manner, these are not explicitly works of social science. They do not specify or define analytic constructs, and a great deal of the information presented is not usable for systematic comparison across time or space.

I conclude with some proposals for realizing some of the best practices in macro-comparative historical research. Ultimately, this involves a tighter and more transparent linkage in the mapping of inherently qualitative observations about the historical record to quantitative classifications.

DEVELOPING NORMS AND STANDARDS FOR HISTORICALLY ORIENTED POLITICAL SCIENCE RESEARCH

The appropriate use of sources and the collection of accurate and unbiased accounts of historical events and processes have been central and longstanding concerns for critics of historically oriented political science. At the extreme, scholars have charged that some studies may be based on "cherry-picked" facts and accounts of

the historical record that are most aligned with a favored hypothesis, biasing the results of those studies (Skocpol & Somers 1980, Lustick 1996, Thies 2002). Quantitative-historical research, particularly in the form of large, cross-national datasets, suffers from distinctive but related pitfalls, and such projects have been charged with a lack of transparency in the coding of cases and lack of attention to appropriate historical context (e.g., Munck & Verkuilen 2002, Bowman et al. 2005, Herrera & Kapur 2007). Although most methodological ink spilled by political scientists has focused on strategies for generating causal inferences, in the absence of reliable and valid data, it is not possible to draw valid inferences about patterns and relationships. Given the centrality of the historical record for political science research, we require greater consensus about how to generate such high-quality data.

Following Gerring's (2001) useful notion of a "criterial framework" for evaluating various aspects of social science practice, I provide one here that is tailored to the collection of historically oriented data. My goal is to identify standards that have been articulated in critiques of historically oriented research, while also highlighting some of the challenges of living up to such standards.

Social scientific analysis of the historical record involves the "disciplining" of unstructured observations of events, outcomes, and specific discourses, converting them into theoretically informed interpretations or classifications. In all cases, the goals of such observational research include accurate portrayals of actual events and discourses, and a consistent application of classification procedures to the record of those events and discourses. Following sound methodological principles is no guarantee of success. Those with a great "feel" for politics and history may sometimes produce better characterizations of the historical record than those who are strictly self-conscious about method. Yet, the promise of a social scientific approach to historical research is that on balance, rigorous investigation, following a set of transparent principles, will generate the best observational data. Here, I identify four of those principles (summarized in **Table 1**): proximity of observations; transparency of citations; certainty of historical record; and attention to valid comparison.

Direct observations: observations recorded by the analyst viewing an event or set of events in real time, through participant observation, ethnography, or interviews with relevant actors

Primary sources: documents or recordings of events and/or attitudes written by individual(s) close to or involved in the subject of investigation

Secondary sources: documents or recordings that describe and/or analyze events and processes based on information that has appeared elsewhere, including in primary sources and/or other secondary sources

Tertiary sources: indexed compilations of information in the form of encyclopedias, factbooks, datasets, or databases

Expert sources: characterizations made by scholars deemed qualified to make judgments and to generate historical narratives based on their longstanding consumption (and production) of direct observations and primary, secondary, and tertiary sources

Proximity of Observations

The central challenge for the collection of historical data—particularly in the characterization of strategies, processes, and events, as compared with strict "count" data (e.g., number of deaths, number of protestors, value of economic transactions)—is to "get the story right." Of course, as John Gerring points out (personal communication), counting is also difficult and subject to related reliability and validity concerns. Notwithstanding, noncount observations are extremely vulnerable to conflicting interpretations, and for larger-scale units, over long periods of time, substantial concerns about what to observe are likely to arise. Which sources can be trusted to provide an accurate account of what transpired at a given place for a given period of time? In order to begin to answer such a question, one must consider the range of source materials used in the generation of qualitative and quantitative data in HIRDs and in historically oriented political science more generally. Sources vary in terms of the proximity of observers—in time and space—to the subjects being described. Ultimately, we are concerned with the quality, or completeness and truthfulness, of the accounts, and this can be evaluated in terms of proximity and the potential for bias within a given context.

It is useful to differentiate among five types of observations in historical research: direct observations, primary sources, secondary sources, tertiary sources, and expert sources. Direct observations are those recorded *by the analyst* when he/she views an event or set of events in real time, through participant observation, ethnography, or interviews with relevant actors. Such observations are generally collected through field research (Wood 2007). Primary sources are those documents or recordings of events and/or attitudes written by an individual

Table 1 A criterial framework for historically oriented political science research

Criterion	Rationale for concern	Assessment	Additional considerations
Proximity of observations	The quality of historical data is only as good as the accuracy of observations. In general, the more proximate the observation to original documents, actors, and events, the less likely observations are to be biased by intermediaries.	To what extent does a research project rely on direct, primary, secondary, or other sources? Are the sources the best ones of any feasible materials for generating data?	Different analytic constructs may be more or less sensitive to biases introduced by observational distance. Primary sources may themselves be biased and/or difficult to interpret.
Transparency of citations	It is impossible to evaluate the accuracy of historical data without knowing the exact source material on which they are based.	Are clear references provided? Are all facts and conclusions specifically referenced? Are original texts made available?	Copyright laws or logistical concerns may prohibit full sharing of original source material.
Certainty of historical record	The density and consistency of historical documentation vary widely. Valid description should clearly report contradictions, strategies for resolution, and reasons for missing cases.	Are contradictory facts or assessments reported and explained? Are missing data identified and explained?	The transparent identification of missing or contradictory data is not at odds with the goal of completeness of data collection.
Attention to valid comparison	Valid descriptive and causal inferences require a fundamental degree of comparability across units. Variable scores ought to have maximal connotation.	Are scores explicitly justified with historical information? Do like scores connote similar phenomena?	In macrohistorical research, there is an inherent heterogeneity across units; assessments of comparability are ultimately subjective.

or individuals close to or involved in the subject of investigation. This is a fairly heterogeneous group of documents and sources that includes the text of important treaties, letters, transcripts of proceedings, and proximate journalistic accounts. Secondary sources are documents or recordings that describe and/or analyze events and processes based on information that has appeared elsewhere, including in primary sources and/or other secondary sources. Tertiary sources are well-indexed compilations of information in the form of encyclopedias, factbooks, datasets, or databases, and HIRDs themselves would be considered tertiary sources. Expert sources are characterizations made by scholars deemed qualified to make judgments and to generate historical narratives based on their longstanding consumption (and production) of the materials described above (see Bowman et al. 2005).

In general, historically oriented scholars tend to put the greatest stock in direct observations and primary sources because their proximity to the events or sentiments under investigation protects them from the layers of distortion subsequent analysts might impose. Scholarship using direct observations and primary sources is often admired because of the sheer effort associated with such work. For example, it is routine practice for political scientists to identify the amount of time they have spent conducting field research, number of sites visited, and/or number of interviews conducted for a particular project as well as the language(s) used in carrying out that research, as markers of the credibility of their work. Primary source materials and detailed historical facts were once the exclusive domain of historians or recognized "country experts" or "area experts" who were willing to invest the time needed to gain access. Today, the proliferation of increased access to a range of government documents, newspaper archives, and a host of facts and primary materials on the internet have provided incredible new opportunities for political scientists to carry out comparative

and historical research across increasingly wide expanses of time and space. Thies (2002) and Trachtenberg (2006) provide guidelines for using primary sources in political science research, making the plea that political scientists ought to rely more heavily on such materials.

But of course, such efforts do not always translate into better data. Nor are they always feasible, and political scientists routinely draw on other types of sources. A study based entirely on primary sources cannot, without further consideration, be deemed more reliable than one based on secondary sources, as some analysts may prove less adept at reading and interpreting the original sources and would do better to rely on secondary sources. Moreover, political actors may purposefully or accidentally misrepresent their role in key outcomes or processes, and researchers are likely to vary in their ability to sort fact from fiction. Working in the broader tradition of comparative-historical analysis, Skocpol (1984, p. 382) argues that a standard requiring extensive use of primary sources by political scientists would make much research infeasible, and that where good secondary sources are available scholars ought to use those. Along these lines, one could argue that a research project based on solid secondary sources—written by scholars who themselves have intimate knowledge of the relevant primary materials and background knowledge and skills to interpret them—is preferable to one based on an overly ambitious, and potentially flawed, attempt to master the primary materials.

Lustick (1996, p. 606) raises substantial concerns about political scientists' reliance on secondary sources without sufficient attention to bias. He recommends more explicit attention to historiography, and the implementation of a consistent strategy that accounts for variance in the work of historians. In practice, however, Lustick's suggestions are exceedingly difficult to implement (Thies 2002, pp. 364–66), and for a great many political science inquiries, a clear historiographical debate is lacking. Thies (2002, pp. 359–64) reviews a set of additional concerns about bias in secondary sources and provides some guidelines for addressing these.

Authorship in data production chains cannot be overlooked (Herrera & Kapur 2007). Moravcsik (1998, p. 10), writing on the integration of Europe, calls for much greater reliance on primary source material and makes the interesting distinction between "hard" and "soft" sources, which he distinguishes according to the likelihood of actors making false statements based on the timing and context of those statements. As a general standard, he ranks the quality of sources from best to worst starting with "hard primary," then "soft primary," then "hard secondary." Moravcsik essentially dismisses "soft secondary" accounts, but this is perhaps because his main focus is on uncovering the specific thoughts and motivations of leaders in that work. Having defined high evidentiary standards for historically oriented political science, Moravcsik's own work was extensively scrutinized and faced strong criticism from a set of scholars who challenged the quality of sources as well as the interpretation of particular texts. For example, Lieshout et al. (2004) examined 221 references in a section on De Gaulle's European policy (Moravcsik 1998, pp. 176–97), and were so critical of the use of source materials that they concluded that the "revisionist" account of De Gaulle's policy could not be supported. This type of exchange suggests the emergence of rising expectations for the quality of direct historical evidence in order to generate conclusions about politics.

Which sources are best? No single standard can be established for all historical research; the utility and appropriateness of sources depend on the nature of the concept under investigation. For example, for the scholar seeking to test a structural theory of democracy, the primary accounts of particular actors, especially ordinary citizens, might be less relevant than secondary and even tertiary data on the timing of economic developments and/or changes in the functioning of a political regime overall. By contrast, for a study focused on the decision making of particular leaders, more proximate observations in the form of those

leaders' letters, direct interviews, and specific statements are necessary (Moravcsik 1998, p. 80; Lieshout et al. 2004, p. 91) because secondary sources may inadvertently make claims about actions and motivations without direct evidence. Whether a scholar can rely on a secondary source's explicit reference to and analysis of a primary source or set of sources depends on the degree to which the secondary author is a trusted scholar who has collected and analyzed source materials in a transparent manner.

Owing to the challenging demands of consistent assessment of source materials across various units, authors of HIRDs may sometimes rely on scholarly experts—credentialed by their skill and familiarity with the relevant language, source, and/or topic—to classify such units in terms of a given conceptual framework. A survey of expert classifications may serve as a reasonable first estimate, but ideally, such expert researchers would be used to generate the specific, identifiable observations needed to classify cases. Although country experts may, for example, claim to "know" the mood of a country or to be able to summarize the quality of political practice, good comparative-historical research practice still requires the documentation of such facts or conclusions. Otherwise, expert classifications may be inconsistent or biased in the same ways that have caused concern about the use of secondary sources (see, for example, Benoit & Laver 2007 on inconsistencies in expert coding of party policy positions). A good practice for HIRD development is to identify the researcher(s) for a particular unit study by name, offering appropriate attribution for research conducted, while creating incentives for high standards of scholarship. Otherwise, claims to "expert" coding ought not to signal any particular confidence in quality of measurement.

In practice, HIRDs use a wide variety of source materials. When evaluating a HIRD for a given area of research, one must consider whether more proximate and less biased materials could have been identified in order to describe and to classify the phenomena under investigation. In general, more proximate observations are considered more desirable, but additional considerations of data quality, including the bias or self-interest of the author, must be taken into account.

Transparency of Citations

A separate but related concern is with the degree to which HIRDs actually reference their data to specific observations and sources that could be verified by other scholars. Such verification provides opportunities to correct careless mistakes, to weigh existing codings in light of new evidence, and to make adjustments calibrated to conceptual and theoretical modifications. Although this might seem to be an obvious minimum requirement of good scholarship, in fact, citation styles vary widely, particularly in research endeavors that attempt to collect a great deal of data across time and space. Readers ought to be able to trace the sources used as the basis for particular characterizations and conclusions. As discussed above, the use of hired "expert" coders should not obviate the need for clear citation because almost any useful characterization needs to be made on the basis of a set of specific observations.

Because political scientists carrying out historical research rely heavily on secondary source materials, it is useful to identify the data and sources embedded therein. For example, our confidence in a HIRD's characterization of a particular variable for a particular country can be substantially heightened if, rather than simply citing a particular author, that author's direct analysis of a representative survey is identified; or if, when a secondary source referenced direct analysis of a particular primary document, the original citation were referenced in a manner that identifies it "as cited in" the relevant secondary source.

An even higher level of transparency can be found when HIRDs actually provide the specific documents or texts that are the basis for the analysis. Moravcsik (2010), for example, has recently proposed that qualitative-historical scholarly articles ought to be published with hyperlinked footnotes, allowing readers to gain

easy access to source materials. Indeed, this is one of the powerful benefits of modern information technology, as scholars can offer others direct access to the materials that undergird their work. In certain cases, full posting of original documents may violate copyright laws, but at the very least, extended quotations are permissible, identifying the location of specific texts. Scholars conducting field research ought to document their findings with transcripts, scanned documents, and other visuals via online HIRDs, allowing their efforts to serve additional scholarly projects in the future. Publishers and legal authorities need to specify which types of postings are permissible under existing copyright laws.

The discussion above suggests that a full accounting of source materials would include some delineation of the nature of the sources and a record of how they were obtained. For purposes of scholarly attribution, it is critical for citations to identify the pathway of observation—e.g., "Constitution of 1812 as cited in Smith 1996, p. 32." Such citation style provides at least two benefits. First, it accurately gives credit to the scholar who expended the effort to track down the primary document, providing incentives for other scholars to make such efforts. Second, if, during the process of verification, mistakes were discovered, it would be easier to identify the sources of such mistakes.

Certainty of the Historical Record

A third concern is with the reporting of levels of (un)certainty surrounding the classification of particular social and political phenomena based on available facts. Most qualitative macrocomparative research relies on a wide variety of sources and goes to great lengths to justify overall classifications in the face of highly heterogeneous source materials. In such work, the citation of multiple sources and discussion of contradictory assessments are routine. Although many qualitative researchers are not always self-conscious in their use of methodological language, the relative uncertainty of the historical record is a form of measurement error that may originate from a number of factors. First, there may simply be a lack of good sources, and for certain times and places, one may have no authoritative basis for characterizing a given variable. Second, there may be conflicting accounts of the same phenomenon related to a variable of interest (for an excellent discussion, see Davenport & Ball 2002). For example, one source might describe a given election as "free and fair," whereas another reports on "irregularities," making it difficult to characterize the quality of a political regime. Third, the researcher may unearth multiple and contradictory phenomena, such as growing rates of intermarriage between ethnic groups but also new examples of interethnic violence, making it difficult to score a variable on the nature of interethnic relations. Finally, additional ambiguities may arise in attempting to classify observations using the rules set out in a classificatory scheme or codebook. For example, should a terrorist attack launched from a single country with the tacit support of government leaders be classified as an interstate dispute? These types of dilemmas are routinely reported in qualitative research but rarely disseminated along with quantitative-historical datasets in a manner that can be traced back to particular cases.

Although the strategy of using multiple coders and estimating intercoder reliability is relevant, such assessments often do not address the potential for errors of the type described above. Correspondence of scores among multiple coders may give the false impression of better measurement in cases where very little actual information has been obtained, whereas the unearthing of additional historical records and facts—i.e., more observations—could generate new ambiguities about classification. For example, if the only fact presented to coders was that elections in a given country-year were "widely regarded as free and fair," multiple coders might easily classify the case as a democracy; but if presented with greater details about the extent of voting irregularities, freedom of the press, restrictions on autonomy, etc., coders might be uncertain about how to aggregate such

facts. The latter example generates more uncertainty about a precise coding but provides a more accurate representation of the case under investigation.

One expects a certain degree of measurement error in all social research, but higher standards for measurement procedures promise to reduce the bias and/or to improve the efficiency of estimates of relationships with other variables. At the very least, the quality of our descriptive and causal inferences can only be fully specified if we have an accounting of such measurement error. In some cases, the manifestation of analytical contradictions associated with classifying historical observations within a given schema may force the scholar to rethink the conceptualization of certain variables and operational coding rules. But perfect measurement is never attainable. In qualitatively oriented research, standard best practice is to highlight such ambiguities and to fully narrate the quality of the historical record and its analytical interpretation. By contrast, most historically oriented replication datasets report "unproblematic" characterizations of the historical record. Best social scientific practice should include the reporting of contradictions and/or general uncertainty of description. Imagine, for example, that a particular variable was being scored on a 1 to 7 scale. It would be useful for the reader to know the extent to which the scholar confidently scored a case as "a 3," as "a 3, but possibly a 2 or a 4," or as "likely a 2 or a 3, but possibly a 4 or even a 5."

Attention to Valid Comparison

As Sartori (1970, pp. 1034–35) famously argued, one of the dangerous pitfalls of the quantification of certain concepts in empirical political science is the potential for "stretching" the meaning of those concepts across units in a manner that severely limits their "connotative precision." Because quantitative studies generally rely on larger sample sizes, scholars may be tempted to use the same classification for too wide a variety of phenomena, limiting our ability to make meaningful descriptive, let alone causal, inferences. For example, to label both a collapsed state *and* a well-disciplined military dictatorship as instances of "nondemocracy" may be counterproductive for many analytic tasks. This concern illustrates the more general challenge of making comparisons across somewhat heterogeneous macro-level units.

A related but distinctive warning about comparability comes from King et al. (1994). They highlight that the prospects of making valid causal inferences through comparative analysis depend heavily on the assumption of unit homogeneity. Social research requires that we make certain analytical choices to meet this assumption, and in practice, the notion that macro-level units such as countries are truly homogeneous remains problematic. Consider, for example, the question of the origins of democracy and the contrasting cases of Sweden and China. This pair of countries could be seen as two of many states in a world of states, and thus meaningfully compared; or they could be seen as fundamentally dissimilar entities—like comparing an ocean liner and a dolphin—for which we ought not to expect similar causal effects in the face of similar treatments. These two cases are unlikely to be juxtaposed in a small-*n* study because of obvious concerns about comparability, but large-*n* quantitative studies routinely include both—employing analytical controls to account for observable differences, but proceeding under the assumption that Sweden and China are intrinsically comparable.

Cross-time comparisons pose similar dilemmas. Whether it is reasonable to expect causal effects to operate in the same manner in the 1890s as in the 1990s depends on the question at hand, and on the ability of the analyst to calibrate different contexts to the theoretical model and measures being employed. If one adopts the more inclusive approach in the creation of a dataset—for example, including more cases and longer periods of time—a transparent rendering ought to take seriously the fundamental concern posed here. In general, small-*n* researchers take great care to justify comparisons, including relevant periodization, whereas quantitative macro-historical

research often prioritizes increased sample sizes, with little explicit hesitation about the prospect of violating assumptions of unit homogeneity or engaging in conceptual stretching.

There are no easy diagnoses of conceptual stretching or unit heterogeneity, but susceptibility to such problems is widespread in both qualitative and quantitative macro-historical research. The validity of comparisons must be justified or explicitly argued when aggregating and analyzing data. Two strategies are routinely used in macro-comparative research to maximize the likelihood of valid comparison. First, the researcher may establish clear and reasonable criteria for the inclusion and exclusion of cases. (Although this is a fairly obvious step, it is another that is not always practiced.) Particularly for the development of multi-country datasets, the goal of maximum country and historical coverage tends to be presented as an intrinsic and self-evident goal, with little regard for the possibility that some or even many countries ought to be excluded owing to fundamental dissimilarities with other cases. Second, and more relevant to the task at hand, the researcher can provide sufficient factual information supporting the ultimate classification of cases to allow users to compare like or contrasting instances of the same analytic phenomenon. As a best practice, the tight and transparent linking of scores to the narrative evidence on which they are based is likely to force scholars to maintain consistency in their classifications, or at least to allow others to judge their consistency across cases.

Contained within the format of HIRDs, as they are defined here, lies great potential for addressing such concerns and potential contradictions. Specifically, the simultaneous presentation of narratives and quantitative mappings provides a ready and transparent check on the consistent use of coding rules for quantification or classification. At the extreme, such narratives can enable early diagnoses of conceptual stretching, and should force scholars to adjust their use of concepts and potentially to eliminate certain cases from the analysis. In addition to the suggestions provided by Sartori (1970), Collier & Levitsky (1997) also offer an extensive framework for addressing such concerns. Achen (2002, p. 446) emphasizes that in quantitative analysis, scholars ought to discard observations "to create a meaningful sample with a unified causal structure."

The simultaneous presentation of narratives and quantitative scores should also serve to discipline qualitative data in a manner that facilitates valid comparative analysis. For example, qualitatively oriented social scientists have not always imposed consistent structures on their reporting of facts; nor have they always clearly described the procedures by which they generate descriptive inferences. The very nature of HIRDs should force scholars to provide explicit analytic interpretations of their qualitative data.

REVIEW OF EXISTING MIXED DATA REPLICATION DATABASES

In this section, I review a range of HIRDs created or substantially informed by political scientists across several substantive research areas. I attempt to apply the criteria developed above as the basis for comparison. Although implicit or explicit causal theoretical frameworks undoubtedly motivated the construction of all of these projects, the data presented in these HIRDs are measures or descriptions of the specific manifestations of analytic constructs, divorced from any hypothesized causes or consequences. As such, I consider these to be conceptualization and measurement projects, and I reflect on the general features of the design of each database that are likely to facilitate or impede scholarly attention to these concerns.

Several HIRDs exist within each of three broadly conceived areas of research: (*a*) regime studies, which characterize, for example, the extent of democratic and/or authoritarian institutions and the nature of leadership selection and replacement; (*b*) conflict studies, which investigate inter- and intrastate war and peace, including the specific treatment of minority groups; and (*c*) gender studies, which describe the nature of policies and/or outcomes associated with the treatment of women.

Tables 2–4 summarize the eight HIRDs identified for these three research areas, providing a framework for comparison. I do not make specific assessments of the ultimate quality of the scores reported for the concepts being measured in any given project. Rather, my intention is to highlight the range of practices used for generating and reporting qualitative and quantitative data so that scholars working in these areas may evaluate the validity of these data. To be clear, I am concerned here only with issues of transparency and the methodological approach to providing access to replication data, not with the accuracy of any particular codings.

Democracy/Regime/Leadership Studies

Few descriptive endeavors have been as central to the study of political science as the characterization of political regimes, particularly at the level of the national state. In many ways, the entire discipline is centered on concerns about the causes and consequences of approximating a democratic ideal. Related qualitative and quantitative research continues to proliferate, with scholars relentlessly retesting established claims—for example, Boix & Stokes (2003) challenging the seemingly definitive findings of Przeworski & Limongi (1997) and Przeworski et al. (2000). Such research requires classification of regime types across time and space. Yet, scholarly disagreement persists about particular classifications.

Munck & Verkuilen (2002) provide an excellent review of several democracy datasets, and of the nine they consider, only two—the Annual Survey of Freedom (Freedom House 2009) and Polity (Marshall et al. 2009)—also include sufficient qualitative data to be considered HIRDs. I refer to the most updated versions of each HIRD as of October 2009. In the case of Polity, this is the Polity IV dataset. Both are extremely well-known, well-established research projects with substantial resources and administration behind them, and each attempts to assess—with slightly different definitions and component features—the degree to which countries have been characterized by political processes that meet the criteria for liberal democracy. In addition, I consider a third HIRD—Archigos: A Data Set of Political Leaders (Goemans et al. 2009)—which is a more recent effort by a small team of scholars to record the manner of entry and exit of national political leaders. All three attempt to classify virtually every country in the world, although some very small countries or territories are not included. What these three share, that other notable "democracy datasets" (e.g., Przeworski et al. 2000, Vanhanen 2000, Brinks & Coppedge 2006) lack, is an extensive and publicly available narrative associated with each country- or leader-case in the dataset. All three rely on a wide range of sources, although only the Annual Survey of Freedom and Archigos seem to use primary/newspaper sources to an extensive degree.

The Annual Survey of Freedom and Polity projects stand out for the scope of quantitative data provided. The Annual Survey of Freedom includes more countries (193 as compared with 163), while Polity includes a much longer time frame (1800–2008 as compared with 1972–2009). Polity contains information on approximately 30 variables, ranging from broadly defined outcomes, such as "level of autocracy," to the date on which a particular regime switched. The Annual Survey of Freedom includes fewer than 10 variables.

Both the Annual Survey of Freedom and Polity report narratives of their regime variables in the form of country reports, and these generally support the quantitative codings. But in both cases, the publicly provided narratives are almost entirely devoid of citations. Although many of the facts contained within such reports are not likely to be widely disputed, others demand greater transparency, including assessments of the degree to which elections were free and fair, the extent of corruption, and press freedoms.

In the case of the Annual Survey of Freedom, the reports and scores are based on "a broad range of sources of information—including foreign and domestic news reports, academic analyses, nongovernmental organizations, think

Table 2 Political regime and transition HIRDs

	Polity IV (Marshall et al. 2009)	Annual Survey of Freedom, country ratings (Freedom House 2009)	Archigos: A Data Set of Political Leaders (Goemans et al. 2009)
Profile			
URL	**http://www.systemicpeace.org/polity/polity4.htm**	**http://www.freedomhouse.org/template.cfm?page=15**	**http://mail.rochester.edu/~hgoemans/data.htm**
Unit of analysis (no. units)	Countries (163)	Countries and territories (193)	Country leaders (188 countries)
Time span	1800–2008	1972–2009	1875–2004
Narratives	Country reports (1 per country)	Country reports (yearly since 2002)	Facts of leadership turnover
Evaluation			
Proximity of observations	Ambiguous, e.g. "multiple historical sources"	Direct observation; primary and secondary sources	Primary, secondary, tertiary, and "expert" judgments
Transparency of citations	None	None	Complete for all facts
Certainty of historical record	No discussion of contradictory observations, but checks on intercoder reliability	No mention of uncertainty	Identifies instances of missing data owing to lack of information
Attention to valid comparison	Country reports are organized around polity component scores, serving as justifications of classification; special authority codes for extraordinary years of foreign intervention or anarchy	Summary scores are reported with country reports, but no explicit justification of components, so difficult to assess validity of comparisons	Owing to narrow focus, not a substantial concern

Table 3 Conflict HIRDs

	Minorities at Risk (MAR) dataset (Center for International Development & Conflict Management 2009)	Uppsala Conflict Data Program (UCDP) database (Uppsala Conflict Data Program 2009)	Alliance Treaties Obligations and Provision (ATOP) database (Leeds 2005)
Profile			
URL	**http://www.cidcm.umd.edu/mar/**	**http://www.pcr.uu.se/research/UCDP/index.htm**	**http://atop.rice.edu/home**
Unit of analysis (no. units)	Minority groups (>280)/flexible	Country (flexible)	Treaties (665)
Time span	1945–2006	1989–2007	1815–2003
Narratives	Group assessments and chronologies	Country overviews and detailed narratives of conflicts	Detailed coding sheets containing specific passages from treaties
Evaluation			
Proximity of observations	Largely secondary; some tertiary and some primary (wire service news reports)	Largely primary (wire service news reports); NGO reports; and "expert" accounts	Largely primary documents (treaties); some newspapers and secondary sources when treaty could not be located
Transparency of citations	References at end of group assessment, but lack of specific citations for each fact	No clear citations	Clear identification of alliance citation
Certainty of historical record	No mention of degree of certainty in historical record; no intercoder reliability checks	Clearly identifies instances where multiple sources point to conflicting characterizations	Highlights possibility of missing observations
Attention to valid comparison	Updated identification of selection bias concerns	Largely clear linkage between quantitative scores and specific historical narratives	Coding sheet format creates consistent, explicit link between narrative and classification

Table 4 Gender HIRDs

	Womanstats (Hudson et al. 2009)	Research Network on Gender Politics and the State (RNGS) Project (McBride et al. 2008)
Profile		
URL	**http://www.womanstats.org/**	**http://libarts.wsu.edu/polisci/rngs/**
Unit of analysis (no. units)	Countries (174)	Policy debates (130)
Time span	c. 1995–present	1970–2001
Narratives	Searchable narrative database	Full edited book volume on each policy area; parsed narratives in online database for each policy area
Evaluation		
Proximity of observations	Primary, secondary, and direct observation; explicitly classified	Primary (original policy debates)
Transparency of citations	Extremely tight link between narrative and source material	Narratives in books are well cited
Certainty of historical record	Allows user to clearly identify conflicting accounts and missing data; not clear how conflicts are addressed in quantitative scales	Intercoder reliability checks conducted and reported
Attention to valid comparison	Summary scales created from narrative data contained in database allows clear assessment	Every observation has parallel quantitative and narrative data; data limited to postindustrial countries

tanks, individual professional contacts, and visits to the region," but there is no presentation in the database of which sources are used for what and why. Although the Annual Survey of Freedom scores are based on a long checklist of components, the country narratives do not systematically describe the assessments of these components, so it would be difficult for any user to assess the linkage between observations and the final classifications.

In a similar manner, the Polity country reports provide no explicit citations. The use of sources for classification is completely opaque in the publicly disseminated web-based version: The codebook refers only to the use of "multiple historical sources for each country, along with. . .a variety of standard sources" (Marshall & Jaggers 2007, p. 17). Discussions of ambiguity and intercoder reliability (Marshall & Jaggers 2007, pp. 5–8) suggest a self-conscious awareness of the challenge of measurement, both in terms of unearthing increasingly better information, allowing for retrospective improvements in the accuracy of accounts, and in the consistent scoring of cases (see also Marshall et al. 2002, pp. 43–44). Although the publicly released data do not identify any specific ambiguities, the country reports are organized around the specific component scores, a practice that allows users to assess the fit between evidence and classification and to assess the calibration of specific scores to codebook definitions.

The Archigos HIRD is a substantially more modest endeavor, which attempts to describe the nature of executive turnover between 1875 and 2004. The transparent citation of each fact provides scholars with the information used to make classifications. One could use the qualitative characterizations as the basis for some case-study analysis, as the HIRD allows one to unearth the proper names and circumstances of specific leaders, but for any given case, this is a fairly limited set of information.

Given the centrality of regime variables to the discipline, there ought to be sufficient intellectual resources available to thoroughly document the specific facts and events that lead to the classification of countries over time. Although the Archigos HIRD is far more limited than Polity and the Annual Survey of Freedom in the range of concepts considered, from the perspective of transparent citation and classification of historical data, it provides a model

for possible improvements to those and other large-scale HIRDs. Given the ubiquity of analyses that use these scores, improvements in documentation would have a big impact on both qualitative and quantitative research.

All three of these HIRDs have opted for highly expansive criteria for case inclusion. However, in terms of concerns about valid comparisons, the Polity HIRD stands out in two respects. Each country report provides a clear summary of the key facts justifying the specific component scores—albeit with a bias toward the present, which makes it difficult to understand the justification for coding changes over time. Moreover, Polity reserves several classifications for special circumstances, such as "foreign interruption," "transition," or "anarchy," and analysts are appropriately warned that country-years so coded should be approached with caution, particularly before one makes comparisons in terms of "degrees" of democracy. By contrast, all independent countries in the Annual Survey of Freedom project are ultimately coded on a single scale. Although the Polity approach may not address all concerns about conceptual stretching, its coding reflects additional attention to valid comparison. Concerns about valid comparison are less problematic in the Archigos project, which is highly focused in its analytic scope.

Conflict Studies: Interstate War, Civil War, and Ethnic Conflict

A second set of HIRDs describes the nature of tensions and patterns of macro-level violence, including large-scale warfare between various social groups and state actors (see Davenport 2007 and Sánchez-Cuenca & de la Calle 2009 for more general reviews of aspects of this literature, with attention to tradeoffs between case-study and quantitative data). Like regime studies, the broad enterprise of conflict studies is central to the discipline, with vast research programs depending on the resolution of a set of fundamental questions: who is in conflict with whom, where, for how long, with what level of intensity, and (when applicable) resolved in what manner? Again, in order for social scientific analysis of the causes and consequences of such conflicts to generate solid conclusions, ambiguity about the identification and classification of events and patterns must be addressed.

Three HIRDs clearly address such concerns: the Minorities at Risk (MAR) database (Center for International Development & Conflict Management 2009), the Uppsala Conflict Data Program (Uppsala Conflict Data Program 2009), and the Alliance Treaties Obligations and Provisions (ATOP) database (Leeds 2005). Although the Correlates of War (COW) collection is largely considered the seminal and central repository for interstate conflict data, the publicly available data are largely in dataset format, with minimal additional source material provided online. Reflecting the scholarly norms described here, the online dataset standards for COW mandate, "Data sources must be clearly identified. Documentation and/or the data set should contain information allowing identification of the source of each newly collected data point. Archival material (e.g., copies of pages from source materials) will be given to the central COW office for permanent archiving." However, I could not identify original source materials for any of the datasets. In a personal communication, the project director (Paul Diehl) confirmed that only limited archives are available, and most of these are currently available onsite and offline. A set of very brief narratives is available for the Militarized Interstate Dispute (MID) dataset, but these provide only a couple of sentences per dispute, and there is no parsing of analytic variables. However, a great deal of source material is likely to be posted soon and may be available as this article goes to press. Finally, the Armed Conflict Database at the International Institute for Security Studies is a subscriber-only database, and I did not consider it in this review.

The MAR database is a seminal project in the field of ethnic conflict. It was initiated by Ted Gurr in 1986, and for most of the history of the project, it has tracked the status of more than 280 minority groups it deems "at

risk," from 1945 until 2006. Such "politically-active ethnic groups" are identified principally by two conditions: the group has experienced systematic discriminatory treatment compared with other groups in society, and the group is the basis for political mobilization. In February 2009, a more limited version of the database with a more expansive definition of minority groups was released, but it is not considered here because most of the available data do not yet reflect the revised definition. When consulted in October 2009, the MAR website reported plans for a future release of a revised dataset. Although research is principally conducted at the group level, this HIRD generates quantitative data in a variety of formats, including at the country-year level, for several dozen variables. That said, although group-year is the most fine-grained level of quantitative data available, it does not appear that there is sufficient information or documentation of the status of MAR groups on a *yearly* basis to treat these as truly independent observations.

For each group, the MAR HIRD provides two distinct qualitative/narrative reports: a group "assessment" and a group "chronology." The assessment provides a set of justifications for several of the quantitative mappings such as "risk status," "group consciousness," and "political discrimination." However, in many instances, the quantitative scores for relevant variables are simply identified in the narrative, without justification, and are presumably based on other (not disseminated) documentation and/or expert assessments. Each report contains a brief bibliography of a few sources, but individual facts and conclusions are not cited specifically. The group chronology provides a set of important facts about the group, marking key events over a long history that extends to available early histories, substantially predating what is available in the quantitative dataset. However, the citation style here is extremely varied—in some cases, particularly for more recent events, facts are documented with specific references to news articles or other sources; but in other places, specific facts and conclusions are entirely undocumented. In general, sources are English-language, with extremely limited citation of specific local sources, except those picked up by the Lexis-Nexis search engine. According to the users' manual (Davenport 2004), no intercoder reliability checks have been conducted. It would not be possible to recreate the range of quantitative variables from the publicly provided qualitative information alone.

The Alliance Treaties Obligations and Provisions (ATOP) HIRD is unique among the HIRDs discussed in this article in that it largely codes very specific alliances for which there exist specific and identifiable texts. The database is thus almost entirely a dissection and systematic cataloging of primary source materials, specifically, military alliance agreements signed between 1815 and 2003. It provides substantial value-added for users as compared with simple access to such primary source materials in that it evaluates those alliances in terms of a set of analytic concerns. Each of the HIRD's 665 alliances is well documented, as publicly available (online) coding sheets provide specific citations of what was actually written in each treaty relevant to the variable under investigation. In certain cases, where the original document could not be found, other primary and some secondary sources are used and cited. Analogous to the Archigos project, the ATOP database provides a clearly documented analysis of a very specific phenomenon, addressing a key concern with a high degree of transparency. Moreover, the ATOP database transparently identifies a set of 90 "candidates" for interstate alliances for which insufficient information was available.

The Uppsala Conflict Data Program (UCDP) provides an extremely well-integrated database of historical facts about organized armed violence and related peacemaking efforts between 1989 and 2007, along with a set of quantitatively coded variables that are easily merged with other quantitative data using the COW data formats. Compared with the other projects reviewed here, the time frame of observation is far more limited—1989 to 2007—but for the case-study researcher, it provides a clear set of narratives concerning various types of conflicts. (The program disseminates strictly

quantitative datasets for more extended time periods.) One can begin an investigation by identifying a country of interest, for which the database provides a summary overview. Substantial narrative information can be identified for various types of conflicts for each country, and additional information can be identified on relevant concerns such as negotiations, third-party involvement, or battle-related deaths. Full texts of peace agreements are also provided. In almost all cases, the linkage between quantitative coding and qualitative reporting of facts is strong and transparent. Database outputs are well organized in terms of variables and years, allowing a high degree of customization for case-study researchers interested in particular dimensions of a conflict.

Despite these virtues, the documentation and citations contained within the narratives are weak. The database overview explains that most information is based on news wire reports gathered through the Factiva database, but there are hardly any citations within the Uppsala HIRD itself. Notwithstanding, it does provide some estimates of potential measurement error in classification. For example, for several citations, the researcher reports on contradictory facts as well as best estimates of the most accurate account. This helpfully reveals that a dated entry is not certain and may be subject to revision.

Gender Studies

Finally, I consider HIRD contributions to the study of gender. Although data on political regimes and group conflict have been in far greater demand by political scientists than data on gender politics and policies, two gender-related databases provide truly outstanding examples of innovative HIRDs. Both the Womanstats database project (Hudson et al. 2009) and the Research Network on Gender Politics and the State (RNGS) project (McBride et al. 2008) are well-integrated presentations of quantitative and qualitative data characterizing the quality of gender relations around the world and, in particular, analytic descriptions of the treatment of women.

According to its homepage (**http://www.womanstats.org/**), Womanstats is motivated by "the link between the security and behavior of states and the situation and security of the women within them." The project has attempted to gather information on more than 250 variables for more than 170 countries. It considers factors such as women's health (e.g., access to health care, female genital mutilation, mental illness); violence (rape and sexual assault, sex trafficking); economic security; security in the state; and security in the family. It identifies a range of quantitative indicators and narrative information of interest for making such characterizations. Moreover, it develops a set of scales based explicitly on the indexed narrative data, allowing an extremely tight calibration between the two.

The project is directed by five principal investigators and draws on a team of undergraduate and graduate research assistants. In addition, it accredits volunteer researchers to add to its living database and includes a web-log (blog) for discussing the data and findings. It provides the foundation for taking advantage of the wealth of scholarly and practitioner knowledge and access to data. It is not clear on what basis accreditation is provided, what interest there has been in contributing to this public good, and/or what credit and responsibility would be accorded to contributors. Nonetheless, this type of semi-open access invites broad opportunities for additions and revisions, while also ensuring some checks on data quality.

Womanstats is the only HIRD reviewed here that allows the user to search for records of narrative data by unit and variable. Every record is tied to a specific source, allowing the user (*a*) to identify multiple records for any given country-variable combination and (*b*) to make distinctions between what the project calls "more generalizable sources," which are laws, statistics and/or authoritative/experts' statements of general fact, and "less generalizable sources," which are predominantly coder inferences, coder comments, nonauthoritative statements/interpretations, extrapolations, and/or personal experiences. It is also possible to

filter results in terms of source types; the sources identified by Womanstats are national government, third-party government, intergovernmental organization, transnational nongovernmental organization, in-country nongovernmental organization, expert interview, journalistic account, non-expert interview, scholarly report, and coder comments. By filtering results according to different types of sources, one could begin to estimate bias in sources in terms of presentation of fact and interpretation. Indeed, it is only with such information that one could begin to formally implement the methodological strategies for avoiding and/or estimating bias suggested by Lustick (1996) and Thies (2002).

It is important to note that the vast majority of sources in Womanstats are English-language and multi-country documents, such as international organization reports. Any scholar interested in the treatment of women across countries would find a substantial starting point in this database, with data richness in the coverage of different aspects of the treatment of women. From an historical perspective, however, the scope is fairly narrow, and there is little in the way of detailed, longitudinal characterizations that would allow the user to identify the origins of particular practices and trends. For a scholarly analysis of either the causes or consequences of such practices and trends, much additional historical research would be required.

The RNGS project offers a different perspective on gender politics, and aims "to document and explain instances of state feminism, that is, those times when institutions inside the state have formed partnerships with women's movement activists to open up the policy making process to include women and women's interests" (**http://libarts.wsu.edu/polisci/rngs/** accessed November 3, 2008). It analyzes a set of policy discussions and policy decisions about various issues with implications for gender equity, and the unit of analysis is the policy debate. The principal investigators of this project describe the unique ways in which their project bridges the "qualitative-quantitative divide" in a paper, "Building a (Data) Bank While Crossing the Bridge: RNGS Strategies to Integrate Qualitative and Quantitative Methods" (McBride & Mazur 2006). They explain how this integrated database was developed through an iterative process of detailing events, processes, and outcomes in narrative terms; classifying them with numerical scores; and then evaluating those measures with tests of intercoder reliability.

The research for this project began with a qualitative phase that involved the authoring of a book on each of five policy areas for 13 postindustrial countries. In a second phase, the authors created numerical mappings of these findings in order to generate a quantitative dataset. Indeed, the ambition of quantifying these data provided a basis for more structured investigation in the edited volumes. Volumes of this type—collections of essays about politics and policy making in a set of countries—often emerge as somewhat disjointed because although similar questions may be posed, divergent strategies for analysis and uncalibrated metrics leave the reader unable to draw general conclusions. By contrast, the RNGS edited volumes are remarkable for the consistency of the structure of analysis across chapters. For example, McBride's (2001) volume *Abortion Politics, Women's Movements, and the Democratic State* contains eleven country chapters written by nine different authors, sandwiched between comparatively-oriented introduction and conclusion chapters. The structure of every chapter is identical, and in each case, the author has closely examined a set of policy debates, largely relying on primary sources, which ultimately become the basis for quantitative mappings. The edited volumes are so consistently organized that one could easily go back to the discussions of the actual texts in order to verify the codings and/or to use that information in a narrative analysis. Such disciplined collaboration would provide an outstanding basis for accumulation of knowledge.

An obvious point of departure for future research would be the simultaneous analysis of quantitative and narrative data from Womanstats and RNGS. For example, are the

status and security of women positively correlated with the level and forms of state feminism, and if so, what is the nature of that relationship? Given the information currently contained in these two databases at present, that question could not be definitively answered without additional data collection and analysis. Nonetheless, what is contained in these HIRDs would provide an extremely valuable starting point, and the information gathered from such an inquiry could make a contribution to these two projects.

CONCLUSION: THE PRACTICE AND PROMISE OF BRIDGING THE QUALITATIVE-QUANTITATIVE DIVIDE

The HIRDs reviewed here make substantial contributions to the practice of political science, and they are all appreciable intellectual public goods within their domains of scholarship and beyond. When they are subjected to emerging standards for the quality of sources, transparency of citations, reporting of certainty in the historical record, and attention to valid comparisons (including the explicit calibration of quantitative scores to qualitative observations), it is clear that much work is to be done. This is particularly striking given that the datasets described here were selected because they were exemplary in their dissemination of complementary qualitative data. I should reiterate that I have not attempted to evaluate the quality of specific scores for any of the data discussed in this review, and I make no claims about their accuracy. My chief concern has been with the publicly described methods for generating these scores.

Notwithstanding these concerns, particularly if more resources were dedicated to the development of user-friendly and flexible software platforms, these pioneering efforts represent possibilities for much greater progress in the accumulation of valid social scientific knowledge. To date, they have facilitated the development of online research communities that extend the reach of contributors and users. They represent the potential for a truly social scientific approach to the analysis of history. Three benefits are particularly striking.

First, HIRDs promise to increase the transparency of the generation of quantitative data. The public provision of flat-file datasets is certainly a step in the direction of transparency in social research, but the application of coding rules to the classification of various social phenomena is often rather opaque. HIRDs encourage the dissemination of the key facts and observations that undergird quantitative mappings, which are often criticized in an unproductive broadside manner. Many coding and classification rules leave open important ambiguities and/or systematically bias classification. Open dissemination of the basis for classification decisions would provide opportunities for scholars to make more specific critiques and productive revisions of existing data, rather than wholesale discrediting of scholarly efforts. Revelations of conceptual stretching might lead to more sound decisions about geographic and temporal reach, such as limiting the domain of comparisons. In this sense, the replication standard can be extended beyond sensitivity to statistical specification, encompassing the careful appraisal of scholars concerned with measurement and the appropriate classification of historical detail.

Too often "random" measurement error is invoked as an excuse for careless measurement in political science, even when it is clearly nonrandom. Beyond the challenges of good description, the quality and precision of statistical estimates of causal effects naturally improve with more precise estimates of the uncertainty of classification. Unfortunately, one consistently striking feature of several of the HIRDs reviewed in this article is the lack of clear citations to sources. Some widely known facts, such as the election of a president or the end of a war through a treaty, may not require citation because they are so widely accepted in the historical record; but a great many facts, including the strategies and actions of particular actors, are frequently conveyed within these HIRDs

without elaboration of how the particular researcher came to know these facts. However, one striking benefit of some HIRDs is direct access to primary texts that remove ambiguity about what is actually being observed. It will never be possible to measure or to classify all phenomena with the same level of certainty, but scholars have an obligation to report the sources of such inconsistency in their dissemination of research. HIRDs facilitate such dissemination.

Second, HIRDs offer opportunities for greater scholarly communication within substantive research areas among scholars using diverse methodological approaches. The relatively open format of HIRDs reflects a better appreciation of the reality of social scientific inquiry: It is an ongoing process, in which mistakes can be corrected and gaps can be filled with effort and further investigation. The RNGS project reveals how rich case-study histories can be developed in a consistent manner to facilitate broader cross-case analyses.

Virtually all historically oriented scholars at some point come up against the problem that in their search for a particular piece of relevant information, no high-quality primary or secondary data are available. Here, the advantage of a well-constructed HIRD should be that lower-quality observations can remain as placeholders until better information becomes available, at which point a database can be updated for future qualitative and quantitative reanalyses. To the extent that case experts are able to use their skills to unearth such information, the fruits of their labor can be more efficiently inserted into quantitative analyses.

Third, HIRDs provide opportunities for scholars to carry out "mixed method" analyses (Lieberman 2005, Gerring 2007). Increasingly, scholars who have estimated certain relationships among variables in statistical approaches to causal inference are choosing also to further test the veracity of those findings using case-study research. For example, Wilson & Butler (2007) highlight that time-series cross-sectional models estimated in political science research—the dominant mode of statistical analysis for the quantitative data described above—turn out to be highly sensitive to alternative specifications, leading to dramatically different results. As a corrective, they conclude that political science is well poised "to pursue a methodological agenda of marrying quantitative and qualitative methods" (Wilson & Butler 2007, p. 122). However, when the selection of cases for such in-depth research is based strictly on theoretical and analytical criteria, this is likely to lead scholars to cases with which they have only minimal familiarity. One approach to this conundrum is to attempt to master entirely new secondary/historical literatures and/or to use general, public-access databases. But such a strategy is likely to be exceptionally time-consuming and/or vulnerable to error. Alternatively, by accumulating well-vetted and theoretically oriented historical narratives of a wide range of cases and for a range of analytic interests within various HIRDs, scholars will have firmer foundations for using case-study data and methods to test the implications of their theories.

Real scholarly progress on concerns of general importance to social scientists requires the development of greater consensus about how we describe key aspects of political life. For example, it will not be possible to evaluate the robustness of theories of democratization if scholars continue to disagree about which countries are democratic. Of course, it would be unreasonable to expect universal agreement on relevant concepts and definitions, but at the very least, scholars ought to be working with the same set of relevant facts when it comes to the analysis of observational-historical data. Area and issue specialists have much to contribute in the contextualization and interpretation of facts, but their input within the social scientific community should be subject to examination, not simply accepted on the basis of informal accreditation. The particulars of relevant contextual concerns and their implications for interpretations need to be laid bare for others to observe and to consider.

As technologies develop, and scholars implement these in new and imaginative ways,

new norms and standards will undoubtedly proliferate. In the meantime, greater adherence to basic standards would certainly increase the quality of quantitative-historical data.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author thanks members of the Comparative Politics seminar at the University of Chicago, John Gerring, Margaret Levi, James Mahoney, Oriana Mastro, Giovanni Cappoccia, and Jaquilyn Waddell Boie for valuable comments and suggestions; and Kristen Harkness for research assistance.

LITERATURE CITED

Achen CH. 2002. Toward a new political methodology: microfoundations and ART. *Annu. Rev. Polit. Sci.* 5:423–50

Achen CH, Snidal D. 1989. Rational deterrence theory and comparative case studies. *World Polit.* 41:144–69

Adcock R, Collier D. 2001. Measurement validity: a shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.* 95:529–47

Bates RH, Greif A, Levi M, Rosenthal J-L, Weingast BR. 1998. *Analytic Narratives*. Princeton, NJ: Princeton Univ. Press

Beck N. 2001. Time-series cross-section data: What have we learned in the past few years? *Annu. Rev. Polit. Sci.* 4:271–93

Benoit K, Laver M. 2007. Estimating party policy positions: comparing expert surveys and hand-coded content analysis. *Elect. Stud.* 26:90–107

Boix C, Stokes SC. 2003. Endogenous democratization. *World Polit.* 55:517–49

Boix C, Stokes SC. 2007. *The Oxford Handbook of Comparative Politics*. Oxford/New York: Oxford Univ. Press

Bowman K, Lehoucq F, Mahoney J. 2005. Measuring political democracy: case expertise, data adequacy, and Central America. *Comp. Polit. Stud.* 38:939–70

Brady HE, Collier D, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Berkeley, CA: Rowman & Littlefield and Berkeley Public Policy Press

Brinks D, Coppedge M. 2006. Diffusion is no illusion: neighbor emulation in the third wave of democracy. *Comp. Polit. Stud.* 39:463–89

Center for International Development and Conflict Management. 2009. *The Minorities At Risk (MAR) Project*. **http://www.cidcm.umd.edu/mar/**

Collier D, Brady HE, Seawright J. 2004. Sources of leverage in causal inference: toward an alternative view of methodology. See Brady & Collier 2004, pp. 229–66

Collier D, Levitsky S. 1997. Democracy with adjectives: conceptual innovation in comparative research. *World Polit.* 49:430–51

Davenport C, Ball P. 2002. Views to a kill: exploring the implications of source selection in the case of Guatemalan state terror, 1977–1995. *J. Confl. Resolut.* 46:427–50

Davenport C. 2004. *Minorities At Risk: Dataset Users Manual 030703*. The Minorities At Risk (MAR) Project, Cent. Int. Dev. Confl. Manage., Univ. Maryland, Baltimore

Davenport C. 2007. State repression and political order. *Annu. Rev. Polit. Sci.* 10:1–23

Freedom House. 2009. *Annual Survey of Freedom, country ratings*. **http://www.freedomhouse.org**

Geddes B. 1990. How the cases you choose affect the answers you get: selection bias in comparative politics. *Polit. Anal.* 2:131–50

Gerring J. 2001. *Social Science Methodology: A Criterial Framework*. Cambridge/New York: Cambridge Univ. Press
Gerring J. 2007. *Case Study Research: Principles and Practices*. New York: Cambridge Univ. Press
Goemans H, Gleditsch K, Chiozza G. 2009. Introducing Archigos: a dataset of political leaders. *J. Peace Res.* 46:269–83
Herrera Y, Kapur D. 2007. Improving data quality: actors, incentives, and capabilities. *Polit. Anal.* 15:365
Herrnson PS. 1995. Replication, verification, secondary analysis, and data collection in political science. *PS: Polit. Sci. Polit.* 28:452–55
Hudson V, Caprioli M, Emmett C, et al. 2009. *WomanStats*. **http://www.womanstats.org**
King G. 1995. Replication, replication. *PS: Polit. Sci. Polit.* 28:444–52
King G, Keohane R, Verba S. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton Univ. Press
Kiser E, Hechter M. 1991. The role of general theory in comparative-historical sociology. *Am. J. Sociol.* 97:1–30
Leeds BA. 2005. *The Alliance Treaty Obligations and Provisions (ATOP) Project*. **http://atop.rice.edu/home**
Lieberman E. 2005. Nested analysis as a mixed-method strategy for comparative research. *Am. Polit. Sci. Rev.* 99:435–52
Lieshout RH, Segers MLL, van der Vleuten AM. 2004. De Gaulle, Moravcsik, and *The Choice for Europe*: soft sources, weak evidence. *J. Cold War Stud.* 6:89–139
Lustick IS. 1996. History, historiography, and political science: multiple historical records and the problem of selection bias. *Am. Polit. Sci. Rev.* 90:605–18
Mahoney J, Goertz G. 2006. A tale of two cultures: contrasting quantitative and qualitative research. *Polit. Anal.* 14:227–49
Mahoney J, Rueschemeyer D. 2003. *Comparative Historical Analysis in the Social Sciences*. Cambridge/New York: Cambridge Univ. Press
Marshall M, Gurr T, Davenport C, Jaggers K. 2002. Polity IV, 1800–1999: comments on Munck and Verkuilen. *Comp. Polit. Stud.* 35:40–45
Marshall MG, Jaggers K. 2007. *Polity IV Project: Dataset Users' Manual*. Cent. Systemic Peace, George Mason Univ.
Marshall M, Jaggers K, Gurr T. 2009. *Polity IV Project: Political Regime Characteristics and Transitions, 1800–2008*. **http://www.systemicpeace.org/polity/polity4.htm**
McBride DE. 2001. *Abortion Politics, Women's Movements, and the Democratic State: a Comparative Study of State Feminism*. Oxford/New York: Oxford Univ. Press
McBride DE, Mazur AG. 2006. *Building a (data) bank while crossing the bridge: RNGS strategies to integrate qualitative and quantitative methods*. Presented at Conf. Br. J. Polit. Sci., London
McBride DE, Mazur AG, Outshoorn J, Lovenduski J, Guadagnini M. 2008. *Research Network on Gender Politics and the State Project*. **http://libarts.wsu.edu/polisci/rngs/**
Moravcsik A. 1998. *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. Ithaca, NY: Cornell Univ. Press
Moravcsik A. 2010. Active footnotes and replicability in qualitative methods. *PS: Polit. Sci. Polit.* In press
Munck GL, Verkuilen J. 2002. Measuring democracy: evaluating alternative indices. *Comp. Polit. Stud.* 35:5–34
Przeworski A, Alvarez ME, Cheibub JA, Limongi F. 2000. *Democracy and Development: Political Institutions and Material Well-Being in the World, 1950–1990*. Cambridge, UK: Cambridge Univ. Press
Przeworski A, Limongi F. 1997. Modernization: theories and facts. *World Polit.* 49:155–83
Sánchez-Cuenca I, de la Calle L. 2009. Domestic terrorism: the hidden side of political violence. *Annu. Rev. Polit. Sci.* 12:31–49
Sartori G. 1970. Concept misformation in comparative politics. *Am. Polit. Sci. Rev.* 64:1033–53
Skocpol T. 1984. Emerging agendas and recurrent themes in historical sociology. In *Vision and Method in Historical Sociology*, ed. T Skocpol, pp. 356–91. New York: Cambridge Univ. Press
Skocpol T, Somers M. 1980. The uses of comparative history in macrosocial inquiry. *Comp. Stud. Soc. Hist.* 22:174–97

Thies CG. 2002. A pragmatic guide to qualitative historical analysis in the study of international relations. *Int. Stud. Perspect.* 3:351–72

Trachtenberg M. 2006. *The Craft of International History: A Guide to Method*. Princeton, NJ: Princeton Univ. Press

Uppsala Conflict Data Program. 2009. *Uppsala Conflict Data Program Database*. **http://www.ucdp.uu.se/database**

Vanhanen T. 2000. A new dataset for measuring democracy, 1810–1998. *J. Peace Res.* 37:251–65

Wilson S, Butler D. 2007. A lot more to do: the promise and peril of panel data in political science. *Polit. Anal.* 15:101–23

Wood EJ. 2007. Field research. See Boix & Stokes 2007, pp. 123–46

Annual Review of
Political Science
Volume 13, 2010

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://polisci.annualreviews.org/