The Structure of Description: Evaluating Descriptive Inferences and Conceptualizations

Marcus Kreuzer

Explanation presumes description. Description explores the who, when, where, and how, and its answers furnish the raw material for theorizing and explaining. This connection between description and allegedly serendipitous exploration contributed to the notion that description is inherently subjective and thus incapable of being evaluated. I challenge this notion of "mere" description. I show that description has a distinct structure that consists of discreet analytical stages facing distinct inferential challenges. The quality of description thus becomes a function of how well it addresses those challenges. I explicate distinct criteria for evaluating how well a describer handles those challenges. I illustrate their utility by applying them to the controversy in the late 1990s between Daniel Goldhagen and Christopher Browning over what explained the willingness of ordinary Germans to kill Jews.

The failure to explain is caused by a failure to describe.[1]

—Benoit Mandelbrot

Mathematician and polymath, 1924–2010

Mandelbrot's admonition to properly describe before setting out to explain may seem startling, especially coming from a world-renowned mathematician, trained in arguably one of the least descriptive disciplines. But his admonition resonates with political scientists doing process tracing, case studies, or comparative historical analysis, and who trade off technical rigor for a more descriptive, exploratory, theory-building mode of analysis.[2] Their work testifies to a recognition that description is something distinct and crucial for generating theoretical insights.[3] Yet for all its appreciation, "mere" description still lives under the shadow of explanation and does so because it lacks evaluative criteria.[4] This paper puts the canard of mere description to rest by demonstrating that description has a clear structure, involves distinct inferential tasks, and makes it possible to ultimately differentiate bad from good description.

The paper is organized into three sections. The first sketches the description conundrum that while political scientists widely agree upon the importance of description, but they disagree over whether and how it can be evaluated. It attributes this conundrum to the conflation of two forms of description, historical and statistical description, that have to be evaluated differently. The second section outlines five key elements of description: finding new facts, organizing them through concepts, selecting evidence from facts, specifying the ontological scope conditions of evidence, and making cross-level inferences. It shows that historical and statistical description share these five elements and that each has its own criteria against which these five steps can be evaluated. In short, I argue that historical description, once it is analytically differentiated from statistical description, can be just as readily assessed.

The third section illustrates the ability of these criteria to discriminate between bad and good historical

*Annotations for Transparent Inquiry (ATI) for this article are available on the qualitative data repository as Kreuzer, Marcus. 2018. "Data for: The structure of description: Evaluating historical description and its role in theorizing". Qualitative Data Repository. https://doi.org/10.5064/F6OAEIDA.

Marcus Kreuzer is Professor of Political Science Department at Villanova University, (Markus.Kreuzer@Villanova.edu). *He works on the formation of party systems in interwar and post-communist Europe, the origins of modern democratic institutions, and qualitative methodology. He is currently writing a book on comparative historical analysis. He would like to thank Stefanie Walter, Giovanni Capoccia, Julia Lynch, Mark Pollack, Orfeo Fieretos, Hillel Soifer, Evan Lieberman, Maria Toyoda, Jennifer Dixon, Kunle Owolabi, and Michael Bernhard for their comments. The four anonymous reviewers were terrific and improved the article beyond recognition. Many thanks for Colin Elman, Diana Kapiszewski, and Sebastian Karcher for spearheading the Annotation for Transparent Inquiry (ATI) initiative that this article uses to increase the transparency of its research judgments. A special thank you goes to Volha Charnysh for her thoughtful and detailed ATI feedback.*

doi:10.1017/S1537592718001068

description by drawing on the well-known Goldhagen controversy. The publication in 1996 of Daniel Goldhagen's *Hitler's Willing Executioners: Ordinary Germans and the Holocaust* set off a heated scholarly and public debate.[5] This debate is interesting because only four years earlier Christopher Browning published *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland* that had tackled the same question posed by Goldhagen.[6] As part of their broader analysis, both scholars were describing how willingly ordinary Germans killed Jews during the Holocaust. Both used the same archival sources but ended up describing the perpetrators' willingness in different ways. Their disagreement triggered an unusually large and intense scholarly debate that evaluated their respective analysis. This debate identified flaws in Goldhagen's analysis that illustrate the usefulness of my proposed criteria. Moreover, the debate concluded that Browning's description was clearly superior to Goldhagen's. It underscores that evaluating historical description is not only possible but also can reach a scholarly consensus and thus belies the claim that description is subjective and relativistic and thus inferior to explanation.

The Description Conundrums

Methodologists agree on the broad contours of description, what motivates it, and how it contributes to social inquiry. Despite this consensus, the effort to assess description faces two challenges. First, constructivists raise a fundamental question about whether facts provide an epistemologically defensible benchmark for evaluating description. They point out that descriptive inferences rely on theory-laden evidence which makes it problematic to evaluate description strictly on their factual basis and without consideration of theoretical presuppositions. Second, methodologists have written only sparsely on how they evaluate description and their limited writings put forth different evaluative criteria. So, the consensus on why we describe is challenged by the question whether description can be evaluated, and if so, how it is to be evaluated. I need to address these two conundrums description faces before I show how to evaluate it.

Consensus on Why We Describe

Description is recognized across disciplines and methodologies as a crucial element of social inquiry. Historians of science discuss its role in the development of modern science,[7] anthropologists link thick description to understanding,[8] sociologists emphasize its centrality in theorizing,[9] and political scientists discuss its importance for concept formation.[10] These discussions treat description in very general terms and associate it with exploring the social world by finding out, just like journalists, *who* the central actors were, *how* they behaved, under *what* circumstances, and *when* and *where* their actions took place.[11] These explorations help to clarify "what the devil is going on around here";[12] to name, abstract, and categorize social occurrences; to discover new dimensions disguised by previous concepts;[13] and ultimately to "establish that the empirical puzzle really exists, that the thing-to-be-explained is there to be explained."[14] Finally, these discussions also recognize that description plays a crucial role in theorizing by helping to re-specify theories and thereby untangle test anomalies.[15] In short, there is broad agreement that description translates factual observations into testable hypotheses and thus connects the empirical complexities of social reality with the technical testing requirements of social inquiry.

First Conundrum: Can Description Be Evaluated?

This broad agreement on the importance of description, however, does not translate into corresponding agreement about how to evaluate it. Constructivists contend that the role played by facts in generating description is influenced by its broader historical, cognitive, professional, economic, political, and theoretical context and thus point out that this broader context raises doubts about the epistemological standing of facts. This constructivist challenge is thought-provoking but fully engaging it would take me too far afield. Browning and Goldhagen's analysis might have been influenced by their religious beliefs, their career stages, or of the political implications of their findings. Such factors undoubtedly can matter but they are too random and subjective to be methodologically relevant. I therefore background all these contexts except for the theoretical one because it has the most direct methodological implications (refer to online Annotation 1).

Thomas Kuhn famously pointed out that observations are inherently theory-laden, by which he meant to indicate that any inference drawn from facts is heavily structured by whatever theoretical foreknowledge a scholar uses to select those facts.[16] He meant to challenge the notion that facts are "particulars isolated from their context and immune from the assumptions of . . . theory, hypothesis, and conjecture." Facts become "evidence that has been gathered in light of—and thus in some sense for—a theory or hypothesis."[17] Kuhn's point raises doubts about the epistemological status of facts and my claim that description can be evaluated in terms of its factual foundations. These doubts require a response.

At a general level, Kuhn's claim is true and impossible to refute because no fact is ever entirely pre-theoretical. But if we look at the particulars of the research process, it is possible to demonstrate that facts are sufficiently autonomous from theory to provide an epistemologically defensible basis for evaluating description. I draw support for this claim from historians of science and methodologists.

Historians of science point out that the epistemological status of facts became more ambiguous as the techniques for scientific inquiry improved (i.e., experiments, statistical inference, quantification) and as theoretical

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

knowledge accumulates.[18] The growing role that theory plays in ladening facts with foreknowledge also diminished the inductive potential of facts. Historians of science thus broadly support Kuhn's claim at a general level. But by placing it in a historical context, they also show that the theory-ladeness of facts is not a fixed given, but varies with the level of theory development and formalization of testing techniques (refer to online Annotation 2).

The etymology of the term "fact" underscores its epistemological autonomy. The term itself has a confusing dual connotation (refer to online Annotation 3). The term was adopted during the scientific revolution to "provide a new epistemological category that made it possible, at least in principle, to distinguish data [i.e., facts] from evidence—i.e., to imagine the pure experience, uncontaminated by inference or interpretation."[19] This pre-theoretic understanding of facts contrasts with its use for designating a theory to be a fact after it has been extensively confirmed. (e.g., evolution is a fact).[20] The word "fact" thus carries the two epistemologically contradictory connotations of being independent from theories or being the theory itself. Furthermore, historians of science point out that changing professional labels reflects the variability of theory-ladeness. Over the course of the scientific revolution, natural philosophers or natural historians became natural scientists, naturalists became biologists and geologists,[21] antiquarians and chroniclers became historians,[22] and astronomers became astrophysicists. These labels denote a shift to a less inductive and more theoretical model of inquiry. The older labels survive today and designate more exploratory modes of inquiry. Overall, historians of science make it clear that, while facts are theory-laden, theories do not pre-determine them and hence do not deny them epistemological standing. Theory and facts are intertwined in a dialectical relationship rather than an unresolvable chicken-and-egg conundrum (refer to online Annotation 4).

Second Conundrum: How to Evaluate Description?

Given that it is epistemologically defensible to evaluate description, we now face the task of finding criteria for such an evaluation. This is a challenging task because little has been written on this subject and even less has been agreed on. In political science, Gary King, Robert Keohane, and Sidney Verba (KKV) and John Gerring provide the most prominent treatments of description, but they say little about how to evaluate it and their writings conflict. I will show that this second conundrum can be resolved by differentiating more carefully between statistical and historical description.

KKV provide one of the few systematic efforts to evaluate descriptive inferences. They point out that description, just like explanation, involves drawing inferences from observable pieces of evidence to broader, unobservable target claims. The who, what, when, where and how of a particular event are often not directly observable and thus have to be inferred from observable, but at best circumstantial, evidence.[23] They identify three criteria for evaluating such descriptive inferences: unbiasedness, efficiency, and consistency. There are two problems with these criteria.

First, KKV offer conflicting definitions of description that conflate its historical and statistical variant and that don't consistently align with their three criteria. In some passages, they employ a qualitative understanding of description as historical or case study-based description.[24] They then define description as collecting facts, which is oddly narrow given that collecting facts is at best a minor part of description.[25] At another point, KKV claim that interpretation is somehow something different from, rather than being part of, description.[26] And a longer section relates description to sorting out systematic and non-systematic factors, which is consistent with statistical but not historical description.[27] These definitional inconsistencies suggest that KKV equate description with statistical description and view historical description as something different, something they loosely associate with interpretation, case studies and non-systematic factors (refer to online Annotation 5).

Second, KKV's tacit equation of all description with statistical description explains why their evaluation criteria follow strictly frequentist logic. This logic is evident in their advice to increase the number of observations to meet three evaluative criteria—unbiasedness, efficiency, and consistency.[28] But they don't spell out how to increase observations involving particularizing, non-standardized, historical evidence. Such evidence will never generate the frequency distributions necessary to apply their three criteria. KKV thus ignore that confidence in inferences does not just follow a frequentist logic, in which only the number of observations matter, but that it also follows an interpretive logic in which the quality of evidence and its ability to discriminate among competing hypotheses are crucial (refer to online Annotation 6).

John Gerring, in turn, is ambivalent about whether description can be evaluated. He articulates detailed criteria for evaluating concepts, which are a key element of description, and uses them to identify the shortcomings of existing democracy indicators.[29] Yet despite evaluating these specific descriptions, he remains skeptical about the ability to evaluate description in general. He contends that "causal inference is still a more highly structured—more 'objective'—enterprise than descriptive inference" and expresses doubt "whether one can say anything at all that pertains to this broad and seemingly incoherent subject."[30] Gerring's verdict is a bit surprising because, unlike KKV, he clearly distinguishes between historical and statistical description. His typology differentiates between a singular version of historical description, which he labels as "particularizing accounts," and

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

four generalizing, statistical forms (e.g., indicators, associations, syntheses, typologies).[31] One therefore has to assume that his skepticism applies to historical and statistical description alike.

The conundrum of how to assess description requires a differentiation between historical and statistical description and a careful extrapolation of their respective inferential tasks. Each of these tasks involves challenges that description faces as well as the criteria used to assess how well the describer solved them.

Elements of Description and Criteria for Their Evaluation

Looked at superficially, historical description involves the mundane exploration of the rudimentary, journalism-like who, when, where, what, and how of a particular event. But looked at more closely, it constitutes a complex analytical process that requires five tasks: finding new facts, conceptualization, selecting evidence, ontological calibration of evidence, and making cross-level inferences. The last four of these tasks require drawing inferences, that is, they involve leveraging the concrete attributes of observable evidence to understand something broader that is not directly unobservable.[32] Such inferences can be causal when they link evidence to unobservable causal claims, or they can be descriptive when they use evidence to describe something unexplored. Description, therefore, can be understood as the analytical product of the inferences drawn from evidence to something that is both unobservable and unexplored.

In historical description, four steps define this inference process and these steps also provide the basis for evaluating the quality of description. First, conceptualization is one goal of historical description. It involves drawing inferences from circumstantial, non-standardized pieces of observable evidence to generalized and standardized attributes of unobservable concepts. Historical description hence has to be evaluated in terms of the validity of such conceptual inferences. Second, the selection of evidence requires drawing inferences about the probative value of the selected evidence against the evidence that was not selected, or not yet discovered. Historical description thus has to be evaluated in terms of the balance between the strength of the supporting evidence and the potential confounding effects of unobserved but potentially available additional evidence. Third, the different temporal and spatial coordinates of historical evidence complicates inferences because the unobservable target claim involves temporal and spatial assumptions that are more homogeneous than those contained in the evidence. The assumptions about the target claim's presumed ontological uniformity hence have to be evaluated against the actual ontological heterogeneity contained in the evidence. Finally, evidence varies in its granularity and requires drawing inferences from evidence observed at one level of analysis to conclusions stipulated at another level. Historical description thus has to be assessed in terms of the validity of such cross-level inferences.

In short, historical description would be impossible were it not for inferences that make a leap from observable evidence to unobservable and unexplored outcomes. Such leaps entail the risk of overlooking confounding factors that ultimately diminish the quality of descriptive inferences. I elaborate on these four inferential steps together with the fifth non-inferential task of finding facts. I contrast them with statistical description to underscore their respective evaluative criteria. I conclude by highlighting the exploratory role that historical description plays and the resulting importance to assess its contributions for theorizing more systematically. As will become apparent, validating historical descriptive inferences depends on marshaling extensive evidence and lengthy interpretations that fit uncomfortably with the word limits and citations practices of many existing journals. This article therefore is accompanied by Annotations for Transparent Inquiry (ATI) which involves a newly evolving citation protocol and technology aimed specifically at giving qualitative scholars additional room to elaborate on their research judgments.[33]

Finding Evidence and Data: Reliability

Description requires new information that can be explored.[34] Tracing the process by which such new information is turned into evidence, which is used in historical description, helps clarify the first evaluation criteria: reliability.

Facts provide the raw material for historical description, but the descriptive potential of facts is limited by their unstructured nature and the challenges they impose on the researcher to bring her foreknowledge to bear to analytically harness them. Historians acknowledge this challenge by distinguishing between facts and evidence. Facts involve the sum total of all the potentially available documentary recordings of historical occurrences. But such facts are disorganized which makes it challenging to find facts relevant for a theory. Finally, finding relevant facts requires sleuthing, language proficiency, familiarity with the organization of archives, knowledge about legal restrictions guiding their access, intuitions of what might have been deliberately omitted or destroyed, and above all, persistence.[35] Evidence, in turn, involves the subset of relevant facts that historians select and use in their description. Lorraine Daston called evidence "facts with significance" because their selection was guided either by theoretical foreknowledge or because a theoretically unladen fact suggests a new theoretical implication.[36] Richard Evans nicely captured the mediating role of theory when he observed that "facts thus precede interpretation conceptually, while interpretation precedes evidence."[37]

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

It is not clear whether statistical description makes an equally sharp distinction between pre-selected and selected information, but the distinction between observation and data captures something similar. Political or economic indicators turn raw social observations into numerical data, that is, observations with significance.[38] Generating numerical data follows a formalized process that uses concepts to standardize evidence and requires technical measurement instruments for turning nominal observations into interval or ratio data. Inter-coder relatability is the key criteria for evaluating the reliability of such converted data. This conversion of observations into data is very theory-laden and constrained by which observations can be mathematically expressed.[39]

Reliability is the criterion for evaluating the process by which facts and observations were found, recorded, and collected. Numerical data has clear reliability criteria in the forms of the explicitness of the coding or sampling protocols.[40] Historians, in turn, use fact checking and source criticism to evaluate the reliability of their evidence. E.H. Carr recommended "to study historians before you study the facts" and thus hinted at the importance of source criticism.[41] Historians recognize that archives and written records are not neutral collections of historical facts; they do not record everything nor do they do so in a disinterested fashion.[42] Source criticism assesses the credibility of the sources used and figures out how much weight can be assigned to each piece of evidence. Fact checking, in turn, assesses whether facts were properly converted into evidence.[43] Such proper conversion requires accurately recording dates, names, locations, correctly translating testimony, or properly citing textual passages.[44] This conversion entails the possibility of factual errors that can be identified by comparing the evidence against the original facts from which it was generated (refer to online Annotation 7).

Conceptualization: Validity

This generation of evidence and data face the challenge that facts and observations are fragmentary, unstructured, and, largely illegible.[45] Carr points out that "facts don't speak for themselves … but only speak when the historian calls on them."[46] Concepts plays a central role in enabling a dialogue between scholars and their facts or observations that makes them legible. Howard Becker states that "without concepts, we don't know where to look, what to look for, and how to recognize what you were looking for when you find it."[47] He also cautions that concepts "are not just ideas, or speculations, or matters of definition. In fact, concepts are empirical generalizations which need to be tested and refined on the basis of empirical research results."[48] Becker suggests here that concepts, while linked to theories, are not fixed and entirely theory-laden; they also are subject to empirical verification and that validity serves as the criterion to evaluate concepts.

Concepts are abstractions and summarize characteristics of a phenomenon that are not directly observable and thus need to be inferred from observable evidence. Validity assesses the quality of these conceptual inferences by asking how accurately a concept summarizes those characteristics arithmetically or figuratively.

In statistical description, concepts are fixed data containers whose validity is assessed in two ways.[49] In the rare instances where the population means are known, the sample mean can be used to assess the validity of a particular concept.[50] Otherwise, concepts are evaluated in terms of their usefulness for generalization. A concept is evaluated in terms of its resonance with existing terminology, the consistency with which it is used across cases, the clarity of its differentiation from neighboring concepts, its utility to describe phenomena across divergent contexts, and how closely it corresponds to the phenomena it purports to describe.[51]

Historians, on the other hand, evaluate concepts in a less formal manner because they don't treat concepts as fixed data containers that are meant to make observations comparable. They treat concepts as loose, flexible proto-concepts that are continuously updated to better describe the relevant facts at hand. This conceptual updating reflects John Lewis Gaddis's point that historical facts need to be made legible by reorganizing and combining them into slightly broader evidentiary categories, some of which might have been suggested by pre-existing concepts. He argues that replicating facts in all their particularities would be of little use because "the reader would drown in detail." What is required instead is "distillation" or "representation."[52] Concepts guide this distillation process but the distillation itself also updates the concepts.[53] Given this different function, historians evaluate conceptual inferences in two ways. First, they argue over the exceptionalism of a concept, that is, whether it is *too unwilling* to generalize. German historians, for example, have long argued over the so-called Sonderweg, that is, whether Germany's path to modernity was unique.[54] Second, historians evaluate concepts on whether they are too theory-laden, *too willing* to generalize, and thus hide facts that should be explored. They focus on the elements that a concept leaves out and thus produces a description that is too particularistic (i.e., under-generalize) or too general (i.e., over-generalize).[55]

Selecting Evidence and Data: Representativeness

Description requires drawing inferences from a subset of selected evidence for the entire evidence that is potentially available. These inferences are judged against different criteria for statistical and historical description. Data is selected through the sampling of observations and is evaluated in terms of the randomness by which the observations were selected and the size of the sample relative to the population. The process for selecting

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

historical evidence is less formalized and hence trickier to evaluate. David Hackett Fischer noted that to analyze history,

> is to be endlessly engaged in a process of selection. No part of the job is more difficult or more important, and yet no part has been studied with less system, or practiced with less method. Many facts are called, but few are consciously chosen, on explicit and rational criteria of factual significance.[56]

Fischer is right that historians lack formal criteria for selecting evidence. But this does not mean that they are not concerned about mis-selected evidence, that is, inconsequential noise or un-selected counter-evidence containing potential confounders.[57] On the contrary, historians employ three serendipity heuristics to reduce the risk of overlooking confounding evidence: they diversify the types of sources consulted, they read history forward to reduce hindsight bias and conceptual reification, and they make their judgments conditional on how many of the available sources were consulted and how exhaustively they were reviewed. Together, these three heuristics follow a Bayesian-like logic that replaces randomization and sampling with subjective probability judgments about the respective likelihood of selecting supporting evidence and overlooking confounding counter-evidence.[58]

First, historians try hard to find new sources that might contain potential counter-evidence not collected by existing archives. This involves diversifying their sources by looking at public and private, as well as domestic and foreign, archives.[59] Historians regularly point to the limitations of their archives as a source of biased inferences because it reduces the probability of having considered potential counter-evidence. Second, historians were aware of the hindsight bias long before psychologists explored it more systematically. Reading history backward carries the risk of a "creeping determinism," which is linking only those dots that causally connect to the outcome to be explained.[60] The same goes for description. Looking back at history through a fixed concept contributes to a creeping conceptual reification and reduces the likelihood of finding new potential counter-evidence.[61] Third, historians occasionally muse that avoiding biased evidence selection is only possible through exhaustive description, that is, exploring and selecting all the relevant facts.[62] They avoid the impracticality of such exhaustive description by estimating the probability of finding counter-evidence in light of the findings of prior scholarship.[63] Carr writes that "historians start with a provisional selection of facts and a provisional interpretation in light of which that selection has been made—by others as well by himself" and then repeats this process.[64] This iterative process helps historians make subjective judgments about how many of the existing facts have already been looked at and how exhaustive their own search consequently ought to be. The three elements guiding the historian's evidence selection—emphasis on diversity rather than frequencies, the extra attention to potentially confounding facts, and the evolving nature of human knowledge—are all elements that figure prominently in Bayesian analysis.[65]

Ontological Calibration: Making Boundary Conditions Transparent

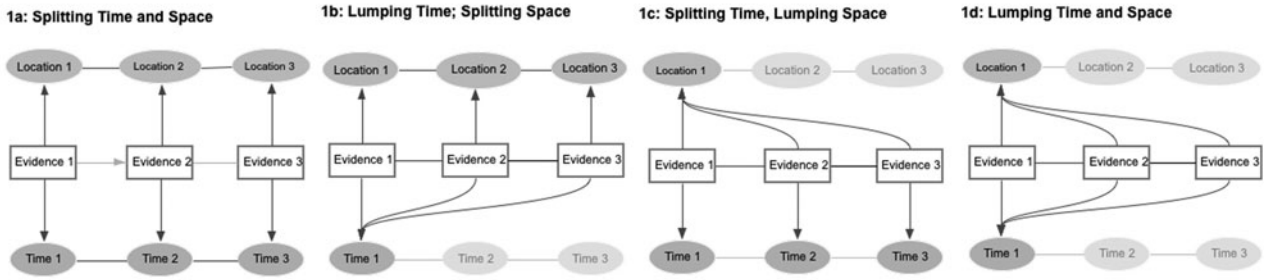
Concepts help discover and organize facts and observations that oftentimes have very distinct chronological and geographic coordinates. Conceptualization thus also requires attention to its ontological calibration. The variable coordinates of pieces of evidence raise the question of whether concepts are subject to historical or geographic boundary conditions and, if so, how clearly those conditions are spelled out. Such boundary conditions are crucial for assessing whether inferences are biased when the pieces of evidence used to support them have different temporal or spatial coordinates. They explicate the ontological assumptions about the uniformity of evidence and thus become the benchmark for assessing how plausible such assumptions are when the spatial and temporal coordinates of individual pieces of evidence vary and the inferences consequently are cross-temporal or cross-spatial. Statistical description barely pays attention to the biases of such cross-temporal and cross-spatial inferences.[66] I therefore explicate the criteria exclusively from the work of historians.

Gaddis argues that historical description makes it necessary to liberate the historian from "the limitations of time and space; the freedom to depart from strict chronology; the license to connect things disconnected in space, and thus to rearrange geography."[67] He further contends that such "re-ordering is again necessary to address the limited physical capabilities of individuals to observe . . . . Events [that] stretch over space and time."[68] Historians thus specify the temporal and geographic reach of their concepts, that is their simultaneity and contiguity. Units of analysis can be single moments, specific events, periods, or pre-specified calendric units (i.e., decades, centuries) Each of these temporal specifications, or what historians would call periodizations, assumes that observed pieces of evidence occurring during this unit of analysis are simultaneous even though in a strictly chronological sense they are not.[69] Or they are assumed to be contiguous at a local, regional, or national—even international—level even though within this geographic confine they took place in different locations. Historians thus detach pieces of evidence from their chronological or geographic context, re-order them, and make them more uniform and hence comparable.

The four panels in Figure 1 present ideal types of cross-temporal and cross-spatial inferences by showing how time and space attributes of evidence can either be lumped to become more simultaneous or contiguous, or can be split to retain their chronological and locational particularities.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

Figure 1
Configuring time and space



The clear boxes in the middle row represent individual pieces of evidence and the convergence of arrows indicates the degree to which their particular spatial and temporal coordinates have been lumped or split.

In panel 1a, each piece of evidence retains its original spatial and temporal coordinates. The analysis makes no ontological simplifications and involves no cross-spatial or cross-temporal inferences. In panel 1b, the spatial evidence is split while geographic evidence is lumped. This situation corresponds to ahistorical, cross-sectional analysis like Theda Skocpol's book on revolutions.[70] Skocpol compares the French, Russian, and Chinese revolutions in their respective geographic contexts but largely filters out their very different locations in time.[71] Panel 1c lumps local particularities into a national story involving a sequence of discreet events. Finally, in panel 1d, three pieces of evidence with distinct geographic and temporal coordinates are lumped together and presumed to take place during the same larger-scale time period and geographic area. The degree of simultaneity and contiguity depends on whether time is calibrated in months, years, or decades, and whether space is calibrated in terms of towns, countries, or regions.

Historians do not have such an explicit criterion for evaluating these ontological simplifications. Just as with timekeeping or maps, the proper calibration depends on what is being represented and for what purpose. The proper level of lumping or splitting has to be assessed relative to the goals of a particular description. However, historians emphasize the importance of being transparent about the boundary conditions of their ontological calibrations because it draws attention to two inference problems: reductionist or exceptionalist fallacies.

Lumping time and space carries a *reductionist risk* because it assumes uniformity in discrete pieces of evidence and thereby overlooks their potential *spatial particularities* and *temporal discontinuities* that may lead to overgeneralized inferences. Spatial particularities make a piece of evidence less uniform because some of its attributes are tied to a locality and therefore are not comparable with pieces of evidence from other localities. In turn, temporal discontinuities make a piece of evidence less uniform in two ways. Either, the piece of evidence is tied to a particular historical period that is characterized by so many one-time contingencies that it is distinct from other periods and therefore cannot be readily compared. History in this instance is "one damn thing after another" (varying attributions). Or, the pieces of evidence interact with each other over time through some learning- or path-dependent process. The discontinuity in this instance could result from the qualitative changes over time, rather than discrete contingencies, that make the evidence different and hence non-comparable.

Splitting time and space, in turn, carries the risk of creating an *exceptionalist fallacy*, that is, the likelihood of overlooking possible inferences. It assumes a lack of evidentiary uniformity and hence misses *potential* commonalities across time and space that might exist among discrete pieces of evidence. Spatial generalities are possible in the presence of powerful diffusion (i.e. technology) or coercive coordination effects that weaken the impact of local particularities and convergence.[72] Temporal continuities, in turn, are possible when evidence is not significantly affected by period effects and remains unchanged. Legacies, for example, refer to pieces of evidence that are comparable across different political regimes.[73] Historians are particularly prone to exceptionalist fallacy because they rarely compare evidence from different countries.[74]

Historical description will always be subject to exceptionalist or reductionist fallacies. The resulting under-generalizations or over-generalizations become problematic only when it can be empirically demonstrated that they omit important confounding factors. And such demonstration requires transparency about boundary conditions of ontological assumptions in the first place.

Inference across Levels

Concepts stipulate not just what constitutes evidence or its boundary conditions but also the unit of analysis at

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

which evidence is collected. This can have important implications for the inferences because oftentimes evidence available for units of analysis are different from units for which the inference is being made. This incongruence between these two units of analysis necessitate so-called *cross-level inferences* which require close attention because they can be subjected to various confounding effects that affect the validity of such inferences as well the quality of the overall description.

Cross-level inferences are well understood in statistical description. Individual-level data is used to draw inferences about groups, regions, or countries, just as group-level data are used to make inferences about individuals. The confounding effects resulting from cross-level inferences are known as the ecological inference problem. Statisticians have developed various techniques for addressing the problem of scaling up from smaller to large units of analysis as well as for scaling down from larger to smaller ones.[75]

Cross-level inferences also pose a challenge for historical description. Gaddis writes that "anytime a historian uses a particular episode to make a general point, scale shifting is taking place: the small, because it is easily described, is used to characterize the large, which may not be." Scaling downward uses evidence from a general category to make an inference of smaller, more particular units and scaling upward uses evidence from particular units to make inferences regarding more general ones.[76] Scaling is essential for abstracting from evidence and generating broader descriptive inferences.

Historians rely on interpretive judgments to address two specific confounding problems of cross-level inferences: the fallacy of composition and fallacy of division. The *fallacy of composition* involves drawing invalid inferences from the actions of individuals to the actions of a group. It would, for example, be unwarranted to infer the patriotism of a platoon solely from the salutes of its individual members on private occasions.[77] The salutes are merely circumstantial evidence that require further evidence to support such an inference. The *fallacy of division* involves drawing hasty inferences from the action of a group about the preferences of its individuals. It would again be unwarranted to assume that all platoon members are patriotic because the platoon marched in a Fourth of July parade. It is important to underscore that cross-level inferences are not automatically invalid, but their validity is conditional on the quality of the accompanying interpretations.

Historians offer interpretations to convince readers that evidence observed at one level supports an inference at another level. These interpretations can be evaluated in four distinct ways. First, how readily do scholars acknowledge the cross-level inferences as well as their magnitude?[78] A personal letter can be the baseline for making cross-level inferences to a family, a group of friends, a police battalion, soldiers in general, or an entire demographic group and thus involve different magnitudes of upscaling. The larger the magnitude the greater becomes the risk of confounding effects. Second, a single piece of evidence invariably provides only circumstantial support for a cross-level inference. An inference therefore can be judged by how many additional pieces of circumstantial evidence are offered.[79] Third, how explicit and detailed the offered interpretation is allows the reader to replicate the reasoning process from the evidence to inferred outcome. And does the explicitness of this inference increase with the magnitude of the inference?[80] Fourth, how readily do scholars address possible alternative interpretations for a cross-level inference? These four elements offer again a quasi-Bayesian alternative to frequentist inferential logic championed by KKV.[81] (refer to online Annotation 8).

Overall, this section demonstrated that description, rather than being subjective and mere, involves distinct analytical steps that impose distinct logistical challenges and can be evaluated. The fact that the evaluative criteria differ for historical and statistical description does not diminish our ability to differentiate good description from bad description, as the next section will show.

The Goldhagen Controversy

Goldhagen and Browning's different descriptions about ordinary Germans' willingness to kill Jews became the basis for their respective explanations. While the ensuing scholarly debate focused on both their descriptive and explanatory inferences, the former played a particularly prominent role. Scholars challenged Goldhagen's description through various means; some reread the trial transcripts, others leveraged their contextual knowledge, and all closely read Browning and Goldhagen to see whether they could replicate all or parts their inferences.[82] This historiographical debate gives the initial impression of disparate judgments coagulating into a general critique. On closer inspection, however, these judgments fall into the five analytical stages of description summarized in Table 1. Before elaborating on these biases, I first discuss why the two authors described in the first place (refer to online Annotation 9).

Why Browning and Goldhagen Describe

Given that sound description is a first step towards a valid explanation, it is important to clarify the sequence between Browning and Goldhagen's description and explanation. Both authors operated within an already sizeable literature on the Holocaust that focused on the roles the German state, concentration camps, and committed Nazis played in killing Jews. This literature, however, left four questions unexplored. Were other Germans besides Nazis involved in killing Jews? Were they forced to kill? How willingly did ordinary Germans

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

Table 1
Overview of descriptive biases

Descriptive Steps	Possible Inference Bias	Actual Biases in Goldhagen
1. Finding facts	Unreliable recording and reporting of evidence	None
2. Conceptualization	Ignoring evidence through mis-conceptualization	Typology only considers killing activities
3. Selecting facts	Cherry-picking of evidence, overlooking counter-evidence	Excludes all self-exculpating evidence
4.a. Simultaneity (cross-temporal inferences)	Lumping → reductionism Splitting → exceptionalism	Lumps 1941-1944 with 1933-1941 and pre-1919. Overlooks confounding effects of time periods
4.b. Contiguity (cross-spatial inferences)	Lumping → reductionism Splitting → exceptionalism	None
5. Cross-level inferences	Fallacy of division of composition	Assumes that what is true for members of military units also is true for individual Germans

participate in such killings? And what might explain the motivations behind such acts? The first three questions involved the who, what, when, where, and how that needed to be answered before the analysis could proceed to the fourth explanatory question—why?

Browning and Goldhagen quickly answered the first two questions about the Nazis' involvement in the killings. Existing scholarship pointed out that roughly two-thirds of Jews were killed outside concentration camps through forced marches, starvation, and above all, mass executions. But there was also strong evidence that many individuals involved in those killings were ordinary Germans rather than ideologically committed Nazis. The question as to whether ordinary Germans were forced to kill Jews required a bit more exploring. But this question was also quickly answered after the records showed that Germans working in the police battalions killing the Jews could recuse themselves without facing a direct penalty.

The central question therefore became how willingly did those Germans participate? It was around this question that much of the controversy pivoted. Browning asks "how did these men first become mass murderers? What happened in the unit when they first killed? What choices, if any did they have, and how did they respond? What happened to the men as the killing stretched on week after week, month after month?" [83] And in a similar vein, Goldhagen generated a "phenomenology of killing," that sought to move the understanding of the perpetrators beyond "mere clinical description of the killing operations" and "convey the horror, the gruesomeness of the events *for the perpetrators*. . . . Blood, bone, and brains were flying about, often landing on the killers, smirching their faces and staining their clothes." [84] Thus, establishing the degree of willingness of those ordinary Germans required descriptions of how they killed Jews and whether the *when* and *where* of those killings interacted with the *how*.

After answering those questions, Browning and Goldhagen proceeded to explain why Germans killed Jews. Their explanations were closely tied to their descriptions of how willingly ordinary Germans killed Jews and how they described the historical and geographic context. Table 2 summarizes the two authors' descriptions about Germans' willingness to kill Jews that they inferred from five observable and, hence, describable activities. These activities include the level of participation, degree of

Table 2
Summary of descriptive inferences

Observable Pieces of Evidence	Goldhagen	Browning
• Level of participation	Very high	High
• Degree of voluntarism	Very high	High
• Psychological harm	None	Common
• Extra-ordinary violence	Common	Unusual
• Delight, bragging	Common	Unusual
Un-observable, descriptive inference →	Very high level of willingness	Moderate level of willingness

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

voluntarism, psychological harm resulting from killings, extra-ordinary violence used in the killings, and bragging about killing Jews.

Goldhagen inferred a very high and Browning a moderate level of willingness. Those descriptive inferences are interesting because they are directly linked to their explanations. Goldhagen attributes the willingness of ordinary Germans to a long-standing, particularly venomous form of eliminationist anti-Semitism. He retraces the long-term historical roots of this anti-Semitism by working backward from the wartime willingness of the members of Police Battalion 101 to anti-Semitic writings in eighteenth- and nineteenth-century Germany. Browning also sees anti-Semitism as an important motivating factor, but he places it in a broader context. He emphasizes the brutalizing effects of the war, fighting on the Eastern Front against the Communist Soviet Union, the role of military peer pressure, and interest in military promotions.

The comparison of Goldhagen and Browning is interesting for three reasons. First, the aforementioned link between their distinct descriptions and different explanations underscores just how consequential description is for theorizing. Second, Browning and Goldhagen used the almost identical archival material for their analysis, thus drastically reducing the likelihood that their diverging descriptions were artifacts of the historical facts they consulted (refer to online Annotation 10). Third, the controversy produced a clear verdict and thus calls into question all historical description as inevitably mere description. The publication of Goldhagen's book and his dismissal of Browning's argument unleashed both a public and scholarly debate that is rarely seen in academia. The scholarly response focused on the consistency of their arguments and re-evaluated much of their evidence. And it produced a near unanimous verdict in favor of Browning after it identified flaws in Goldhagen's argument and, particularly, in how he described Germans' willingness to kill Jews. The following review of this verdict demonstrates that historical description can be evaluated in terms of the five proposed criteria just as rigorously as statistical description.

Finding Evidence

Historians evaluate the reliability of their facts through source criticism and fact checking. Browning and Goldhagen address the reliability of their sources in considerable detail which might explain why historians did not raise any significant questions.[85] Both are cognizant that their facts were generated in the early 1960s, twenty years after the actual events took place; through court testimony summarized by investigators rather than verbatim transcripts; and were potentially shaped by the questions posed and the omission of self-incriminating evidence.[86] Goldhagen and Browning also carefully weigh the biases that might arise from these sources, spell out how they went about assessing their reliability, and why they had confidence in them. This, together with the fact that both used the same sources, explains why the reliability of their sources wasn't an issue.[87]

Goldhagen's thesis was subjected to thorough fact-checking, and fact checkers found factual errors, contestable translations, and inaccurate quotations.[88] But they overlooked that such errors are an inescapable part of research and that ultimately what matters is whether they are systematic or not. And it is this systematic quality that Goldhagen's critics failed to adequately document. His errors therefore are the random errors that every scholar commits and were inconsequential for the quality of his descriptive inferences. They probably ended up attracting attention because they are "easier" to verify than the other elements of description.

Conceptualization

Browning and Goldhagen's conceptualizations illustrate the difference between fixed concepts with limited exploratory potential and looser proto-concepts that are updated in light of new evidence.

Browning employs an informal categorization that reflects the type of killings undertaken by the Police Battalion. He distinguishes between indirect killing activities (i.e., rounding up Jews in ghettos, clearing ghettos, deporting them to concentration camps) and direct killing activities (i.e., mass execution, hunting down and executing Jews in hiding). For each of these activities, he organizes evidence relating to the number of Jews involved, how willingly perpetrators participated, the levels of perpetrators' brutality, and their emotional reactions to the killings. This analytical scheme operates at a low level of abstraction and thus captures the relevant nuances of the perpetrators' actions.

By contrast, Goldhagen uses a fixed classificatory scheme that backgrounds important aspects of the perpetrators conduct. He reduces the police officers' activities to a two-by-two table in which the rows differentiate between killing activities that were ordered by the state and those initiated by individuals. The columns distinguish between killings that were cruel—in the sense of involving brutality above and beyond executing Jews at gunpoint—or not.[89] This scheme conceives all perpetrators as differing only on whether they killed following orders or not, and whether they killed with unnecessary cruelty or not. It silences all activities that were indirectly related to killing Jews (i.e., deportations, clearing of ghettos), efforts of individuals to shirk or formally avoid having to kill Jews, any observations about emotional distress soldiers might have experienced after killing Jews, and any acts of kindness directed toward Jews. The omission of acts of kindness turns out to be immaterial because none were recorded in the testimony.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

However, the omission of actions that revealed ambivalence, unwillingness, and even refusal to kill Jews produces a more monochromatic representation of the perpetrators' willingness. Just like with selectivity, such omissions leave out complexities and make the behavior of Germans more uniform. From this more uniform evidentiary basis, it becomes easier to draw inferences about an extremely high level of willingness.

Ultimately, Goldhagen's conceptualization closely reflects his two key theoretical propositions: Germans' were afflicted by an eliminationist anti-Semitism and that this anti-Semitism was unique to Germans. His classificatory scheme, almost by definition, sees only evidence that fits those two propositions, and makes it impossible to update the concept in light of new evidence. It is ultimately so theory-laden that it makes it impossible to observe variations within German anti-Semitism, as well as to compare it cross-nationally (refer to online Annotation 11). By contrast, Browning uses a much looser set of categories that are less theory-laden and exceptionalist. It produces a more distinct description of the perpetrators' actions, which leads Browning to infer a lower level of willingness.[90]

Selecting Evidence

Goldhagen's theory-laden and exceptionalist analytical frame was not the only factor biasing his selection of evidence. He also stipulated rules for admitting evidence that systematically excluded disconfirming evidence and left out contextual information that altered the probative value of the selected evidence.

Critics trace Goldhagen's evidentiary cherry-picking to his decision to categorically "discount all self-exculpating testimony that find no corroboration from other sources" because to "accept the perpetrators' self-exonerations without corroborating evidence is to guarantee that one will be led down many false paths, paths that preclude one from ever finding one's way back to the truth."[91] Goldhagen defends this decision by arguing that the testimony of Police Battalion 101 was given by the perpetrators and thus inherently biased. Browning is also concerned about such biases, but rather than excluding all evidence, he assesses each piece on a case-by-case basis.[92] Goldhagen's blanket dismissal of large parts of evidence severely skewed the facts he considered for potential evidence. It left him with only two potential sources of evidence: evidence from individuals who did not participate in killings, and therefore could be honest, or evidence of particularly zealously anti-Semitic Germans who were unrepentant and did not care about legal consequences. The first group was virtually non-existent and the second group of zealots was relatively small. Goldhagen therefore excludes so much testimony that it leaves "only a residue of testimony compatible with his hypothesis, and the conclusions are for all practical purposes predetermined."[93]

Goldhagen's critics also point to additional forms of cherry picking. They accuse him of mis-citing passages from scholars to fit his analysis (refer to online Annotation 12). And they point out that he frequently leaves out contextual details in order to increase the confirmatory weight of his evidence. They contend that such streamlining of evidence is systematic rather than random because it leaves out confounding evidence about Germans' high degree of willingness[94] (refer to online Annotation 13). Interestingly enough, Goldhagen's propensity to cherry-pick evidence is also evident in his responses to his critics which elide their central criticisms (refer to online Annotation 14).

Ontological Calibration

Critics paid more attention to Goldhagen's lumping of time than his lumping of space (refer to online Annotation 15). They pointed out that he lumps time by combining prewar and war-time events, as well as treating events during different stages of the war as simultaneous. Each lumping biases his descriptive inferences. The first does so by treating two time periods as qualitatively uniform, when they were not, and the second does so by overlooking important interactions between sequential war-time events.

Goldhagen lumps evidence from the pre-war and war-time periods, arguing that the latter had no significant effect on Germans' willingness to kill Jews; he treats the two periods as qualitatively equivalent and relegates any particularities that distinguish the war-time from the peace-time period to inconsequential background noise. To Goldhagen, the war only mattered to the extent that it gave Germans an opportunity to act on their pre-existing willingness to kill Jews, but it had no confounding effect on their willingness.[95] His critics question this contention. The historian Dirk Moses contends that the extreme circumstances of the war

> are not the occasion for the release of pre-existing preferences, but the occasion for the development of new ones. Christian-bourgeois norms were not just moral inhibitions preventing the expression of a latent, genocidal anti-Semitism: they were a qualitatively different preference structure altogether. The Nazis knew that their anti-Semitism was not the source of their popularity, and it worried them. It is no surprise that they endeavored to keep secret the details of the "Final Solution".[96]

Browning concurs and carefully splits events occurring before and during the war. He points out that "nothing helped Nazis to wage a race war so much as the war itself. In wartime, when it was all too usual to exclude the enemy from the community of human obligation, it was also all too easy to subsume the Jews into the 'image of the enemy', or *Feindbild.*"[97] This difference in their splitting and lumping is reflected in the overall structure of their books. Browning's individual chapters are devoted to a single location and discrete point in time, whereas

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

Goldhagen's chapter are much more prone to lumping together time and space (refer to online Annotation 16).

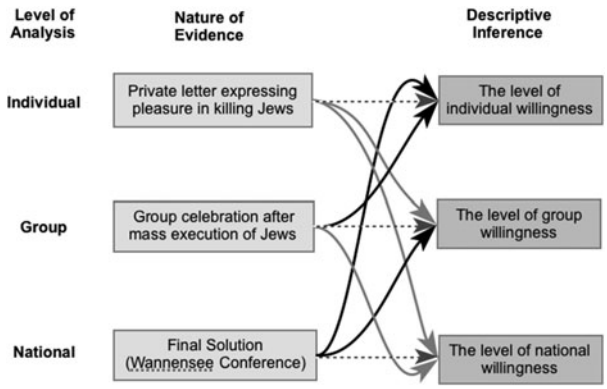
Goldhagen also lumps together events from different stages of the war and thereby misses temporal dynamics and learning effects related to the unfolding of the war itself. His lumping is premised on the assumption that the massacres that occurred over the war years were largely independent of each other and that a perpetrator's action in one massacre did not affect his actions in a subsequent one. Browning questions this independence assumption after he found that the willingness of members of the police battalion to kill and the number of eager killers increased over time.[98] He points to particular feedback mechanisms through which the massacres became interdependent. He argues that the killings themselves, together with the effects of the war, had a brutalizing effect on the members of the police battalion and explains their increased tolerance for killing Jews.[99] Browning also points out that officers learned to reduce the psychological costs of killing Jews after the first mass executions in Józefów in early 1942. They began to recruit SS-trained non-German auxiliaries from Soviet territories for the mass killings and therefore could reduce the frequency with which Germans had to kill Jews. They were then assigned to more regular tasks like clearing the ghetto or supervising deportations that did not involve directly killing Jews.[100] Therefore, to Browning, these temporal effects have a confounding effect that needs to be factored in when drawing inferences from actions to levels of willingness. To properly capture these confounding effects, the actions have to be split into fine-grained, temporally sequenced events that permit observation of their interdependencies.

Overall, Browning's splitting time and space sheds interesting light on why his title ends up referring to the members of Police Battalion 101 as *ordinary men.* To Browning, the battalion members were ordinary men influenced by extra-ordinary circumstances, and so their behavior could be observed by non-German ordinary men acting under similarly extra-ordinary circumstances. Goldhagen's lumping of time and space allowed him to characterize the battalion members in his title as *ordinary Germans* whose behavior was less shaped by the war-time circumstances and more by long-term German-specific eliminationist anti-Semitism. Ironically then, Browning's splitting and contextualization makes his findings more generalizable, while Goldhagen's lumping makes his more exceptionalist.

Cross-Level Inference

Figure 2 shows the cross-level inferences challenges that both authors faced when having to match the scale of their evidence with that of their unit of analysis. It lists the three units of analysis—individual, group, national—used most frequently by Browning and Goldhagen and identifies the

Figure 2
Cross-level inferences



corresponding pieces of evidence. It shows three possible cross-level inferences. The first involves the absence of cross-level inferences if evidence and inference occur at the same level of analysis (e.g., dotted connectors). The second refers to up-scaling by using individual- or group-level evidence to make inference for a higher level of analysis (e.g., grey connectors). Third, downscaling involves using national- or group-level evidence to make descriptive inferences about lower units of analysis (e.g., black connectors).

Browning and Goldhagen's cross-level inferences vary in small but important ways. Both anchor their analysis in the group activities of Police Battalion 101 and other comparable battalions. But Browning is more circumscribed than Goldhagen in his cross-level inferences and provides a more detailed reasoning (i.e., interpretation) when he down-scales and up-scales.

Browning makes very few cross-level inferences since most of his inferences occur at the level of his evidence. When he scales, it involves a modest down-scaling by drawing inferences from the observable group actions to the willingness of individual Germans to kill Jews in similar front-line contexts. Browning is reluctant to upscale and to draw inferences from the police battalion to all wartime Germans or their longer-term anti-Semitism. Such upscaling occurs mostly in his concluding chapter, is highly qualified, and thus cognizant of potential confounding effects. The inferences he draws from the police battalion's action for individual Germans illustrates his attention to confounding effects in cross-level inferences. Like Goldhagen, Browning points out that the policemen could have recused themselves from killing Jews without penalties. He further points out that distressingly few policemen availed themselves of this option.[101] The authors draw different inferences from this evidence. Goldhagen asserts that soldiers had full agency and that

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

broader inferences could be drawn from their individual behavior for Germans in general. By contrast, Browning qualifies this inference by pointing to peer pressure as a confounding factor that mediated actors' private preferences. He therefore questions Goldhagen's cross-level inferences that private motivations revealed by the behavior on the front are a valid predictor for the motivations of ordinary Germans not serving active military units or police battalions. Others have pointed to the confounding effects of the German state. The Nazi regime disseminated extensive anti-Semitic propaganda and imposed considerable costs on political dissent.[102]

Goldhagen is quite up front about his upscaling, claiming that his analysis of the actions of the Police Battalion is "intended to do double analytical duty. This should permit the motivations of perpetrators in those particular institutions to be uncovered, and also allows for generalizing both to the perpetrators as a group and to the second target group of this study, the German people."[103] He further states that the "conclusions drawn about the overall character of the [police battalion] members' actions can, indeed must be, generalized to *the German people in general*. What these *ordinary* Germans did also could have been expected of other *ordinary* Germans."[104] These bold cross-level inferences without adequate interpretations led his critics to accuse him of committing the fallacy of composition (refer to online Annotation 17).

Description and Theory Development

I have made three points: historical description has a distinct structure; this structure contains inferential tasks sufficiently discreet that they can be assessed; and the Goldhagen controversy showed that there is a direct connection between the quality of description and explanation. Since the core of this paper extensively addressed the first two points, I conclude by exploring the connection between description, explanation, and ultimately theory development.

Skeptics might contend that little is to be learned from the Goldhagen controversy because it amounts to little more than a disciplinary turf battle. The skeptics are right that it pits Browning and his fellow historians against the lonely political scientist Goldhagen, and that he was treated no better than many other trespassing social scientists. Such skeptics, however, overlook three important points that, once rebutted, make clearer the broader implications of this controversy.

First, Browning and Goldhagen's disciplinary differences are less relevant than the commonalities of their research question and evidence. Goldhagen used the Holocaust to explore a new dimension of this historical event and not to test theories on genocides. And ironically enough, he, the allegedly generalizing political scientist, produced an explanation so exceptionalist that it even irritated historians (refer to online Annotation 18).

Second, historical description is done by historians and political scientists alike; historians value it to get to the bottom of a particular event, political scientists prize it to get to the bottom of theoretical flaws. Historical description simply is an irreplaceable element of a broader, generalizing, and hypothesis-testing social science.[105] In political science, it is essential for fact checking,[106] validating natural experiments,[107] properly specifying causal mechanisms,[108] or making theoretical sense of testing anomalies. Goldhagen's descriptive flaws therefore have broader methodological implications because getting historical description right matters to political scientists just as much as it does to historians.

Third, cognitive mindsets might have mattered more in the controversy than disciplinary differences in shaping the quality of description. The structure and tone of Goldhagen's analysis epitomizes what Isaiah Berlin famously referred to as a hedgehog-like mindset that approaches analytical tasks with one big, bold, and fixed idea.[109] Goldhagen's opening chapters frontload his bold "no German, no Holocaust" thesis as superior to prior explanations.[110] The subsequent chapters present copious supporting evidence. By contrast, Browning represents a more fox-like mindset that seeks a close and intimate dialogue between different smaller ideas and evidence in order to update the prior knowledge. He begins his analysis with an order of the Police Battalion 101 to execute hundreds of Jews in July 1942, which marked also the beginning of the eleven most deadly months during which over half the Jews were killed who perished during the Holocaust.[111] His subsequent chapters chronicle additional executions over those leven months. It is only in his concluding chapter that Browning draws on broader sociological, psychological, and historical research to explain why these ordinary Germans killed Jews so willingly. This difference in Goldhagen and Browning's cognitive mindsets illustrates a key point made at the beginning of this paper, that description will only fully realize its exploratory and theory-generating promise when it is not overly constrained by cognitive, theoretical, and epistemological priors. It is the relative absence of such priors among historians that makes historical description so crucial for formulating and refining theories. This untethering of description from such priors and its upgrading from "mere" to systematic will make it easier to address Christie Aschwand's cleverly stated conundrum that "it is easy to get results but difficult to get answers."[112] Greater attention to the quality of description will help in sorting out whether tests produce mere results as opposed to genuine answers.

Notes

1 Cited in Hagstrom 2013, 85.
2 George and Bennett 2004.
3 Gerring 2012a, 50–80; King, Keohane, and Verba 1994.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

4 Gerring 2012a, 110–12.
5 Goldhagen 1996a.
6 Browning 1993.
7 Ogilvie 2008; Daston 2011.
8 Duneier 2011; Geertz 1973.
9 Swedberg 2014; Rueschemeyer 2009.
10 Sartori 1970; Collier and Levitsky 1997.
11 Gaddis 2002, 35–53; Gerring 2012a, 107–9; Becker 2007; King, Keohane, and Verba 1994, 33–35; Howard 2017, 36–38.
12 Abraham Kaplan cited in Gerring 2012b, 1.
13 Becker 1998, 109–46; Swedberg 2012, 14–17.
14 Abbott 2016, 198.
15 Abbott 2004; Becker 1998; Swedberg 2012.
16 Kuhn 1962; He borrowed this idea from Ludvik Fleck 1935.
17 Poovey 1998, 1.
18 Daston 2011; Wootton 2015, 38–430; Poovey 1998.
19 Daston 2001, 13.
20 Wootton 2015, 255; Poovey 1998, xii.
21 Daston 2011.
22 Rosenberg and Grafton 2010.
23 King, Keohane, and Verba 1994, 42–43.
24 Ibid. 44–45.
25 Ibid. 34.
26 Ibid. 36–42.
27 Ibid. 53–56.
28 Ibid. 63–74.
29 Coppedge et al. 2011; Gerring 2012a, 116–30.
30 2012b, 722, 730–40.
31 Ibid. 723–27.
32 King, Keohane, and Verba 1994, 46; Becker 1998, 78.
33 The Annotations for this article are available on the qualitative data repository under Kreuzer 201 at. https://doi.org/10.5064/F6OAEIDA. The specific ATI protocol for this article is elaborated in the Data Overview link listed at the beginning of the ATI annotations. More information on ATI can be found at the Qualitative Data Repository.
34 Swedberg 2012, 9–14.
35 Arnold 2000; Markoff 2002; Elman and Kapiszewski 2014.
36 Daston 1991, 94.
37 Ibid.; Evans 1997, 66.
38 Historians of science have long retraced the genealogy of many of our modern-day indicators; Morgan 2011; Porter 1986; Daston and Lunbeck 2011. For the background of more recent indicator, see Coppedge et al. 2011.
39 Hand 2016; Morgan 2011.
40 Lupia and Elman 2014.
41 Carr 1961, 26.
42 Grafton 1999, 34–62; Howell and Prevenier 2001, 60–69; Markoff 2002; Lustick 1996b; Trachtenberg 2006, 140–69.
43 Howell and Prevenier 2001, 43–60; Lustick 1996a; Howard 2017, 146–50.
44 Graves 2016, 82–112.
45 Gaddis 2002, 6–7, 131–36.
46 Carr 1961, 9.
47 Becker 1998, 110.
48 Ibid. 138.
49 Ragin 1994, 74.
50 Gerring and Christenson 2017, 50–51.
51 Gerring 2012a, 116–30.
52 Gaddis 2002, 32.
53 Morgan 2011, 303; Mahoney 2004, 93.
54 Blackbourn and Eley 1984; Kreuzer 2003a.
55 Becker 1998, 122–24.
56 Fischer 1970, 65.
57 Becker 1998, 65–90.
58 Trachtenberg points out that such judgments are just as central in natural sciences as in history. The difference is that historians are more willing to acknowledge their role; Trachtenberg 2006, 14–27.
59 Bennett and Checkel 2014, 27–28; Braudel 1980, 40–48; Howard 2017, 154–57.
60 Fischhoff 1982.
61 Capoccia and Ziblatt 2010.
62 McCullagh 2000, 42; Becker 1998, 76–79; Hurst 1981.
63 Becker discusses various ways to engage existing scholarship to assess the degree of overlooked evidence; Becker 1998, 76–90.
64 Carr 1961, 35.
65 Kreuzer n.d.; Carrier 2012.
66 W. Sewell 1996, 258–61; Hall 2003; Gerring 2012a, 63–65; Knapp 1984, 34–36.
67 Gaddis 2002, 20.
68 Ibid. 25.
69 Zerubavel 2003, 40; Bartolini 1993, 148.
70 Skocpol 1979.
71 W. Sewell 1996.
72 See debate between Ledford 2003; Sperber 2003; Kreuzer 2003b.
73 Wittenberg 2013; Pierson 2003.
74 Sewell 1967, 210–11.
75 King 2013.
76 Gaddis 2002, 25.
77 Fischer 1970, 65–67.
78 Ibid. 62.
79 Becker 1998, 80–82.
80 Fischer 1970, 62–63.
81 Kreuzer 2016; Bennett and Checkel 2014; Zaks 2017.
82 Goldhagen's work has been criticized in book reviews and articles too numerous to be cited fully here. Some of the most important critiques have been compiled in three edited volumes. Heil and Erb 1998; Shandley 1998; Eley 2000;

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

Goldhagen also responded to his critics: 1992; 1996b; 1997.
83 Browning 1993, 37.
84 Goldhagen 1996a, 22.
85 Browning 1993, xxvi–xxxiii, 147–58; Goldhagen 1996a, 463–68, 598–601.
86 Birn 1997, 196.
87 Birn 1997 questions the reliability of Goldhagen's sources. But her criticism is vague and does not point to specific problems.
88 Ibid. 199, 206–07; Browning 1993, 214; Pohl 1997; Stern 1996, 130 n1.
89 Goldhagen 1996a, 17, 376.
90 Browning 1993, 216. Goldhagen harshly criticizes Browning for his lack of conceptual rigor and argues that the additional categories that Browning uses lack any evidentiary foundation; Goldhagen 1992. However, he does not support this criticism empirically, which diminishes its validity.
91 Goldhagen 1996a, 466–68.
92 Browning 1993, 148–58.
93 Ibid. 211.
94 Birn 1997, 196; Blaschke 1998, 71–74; Browning 1993, 211.
95 Goldhagen 1996a, 140–43, 160–64.
96 Moses 1998, 216.
97 Browning 1993, 186; See also Pohl 1997, 40; Mahoney and Ellsberg 1999, 431.
98 Browning 1993, 116, 215.
99 Ibid. 159–62.
100 Ibid. 77. For other lumping problems, see Pohl 1997, 33–34, 40.
101 Ibid. 170.
102 Birn 1997, 197–98, 202-03, 211–12; Moses 1998, 215.
103 Goldhagen 1996a, 464.
104 Ibid. 40, emphasis original.
105 Trachtenberg makes this in a particularly compelling fashion. 2006, 14–29.
106 Lieshout, Segers, and Vleuten 2004; Kreuzer 2010; Parsons 2013; Moravcsik 2013.
107 Kocher and Monteiro 2016.
108 Haggard and Kaufman 2012; Rosato 2003; Narang and Nelson 2009.
109 The connection between these mindsets and the quality of analysis is explored more systematically in Tetlock 2005; see also Silver 2012.
110 Goldhagen 1996a, 6.
111 Browning 1993, XV.
112 Aschwande 2015.

References

Abbott, Andrew. 2004. *Methods of Discovery: Heuristics for the Social Sciences*. New York: W.W. Norton.

———. 2016. *Digital Paper: A Manual for Research and Writing with Library and Internet Materials*. Chicago: University of Chicago Press.

Arnold, John. 2000. *History: A Very Short Introduction*. New York: Oxford University Press.

Aschwande, Christie. 2015. "Science Isn't Broken." *FiveThirtyEight*. Available at http://fivethirtyeight.com/features/science-isnt-broken/, accessed August 20, 2015.

Bartolini, Stefano. 1993. "On Time and Comparative Research." *Journal of Theoretical Politics* 5(2): 131–67.

Becker, Howard. 1998. *Tricks of the Trade*. Chicago: University of Chicago Press.

———. 2007. *Telling about Society*. Chicago: University of Chicago Press.

Bennett, Andrew and Jeffrey Checkel. 2014. "Process Tracing: From Philosophical Roots to Best Practices." In *Process Tracing*, ed. Andrew Bennett and Jeffrey Checkel. New York: Cambridge University Press.

Birn, Ruth Bettina. 1997. "Revising the Holocaust." *Historical Journal* 40(1): 195–215.

Blackbourn, David and Geoff Eley. 1984. *The Peculiarities of German History*. Oxford: Oxford University Press.

Blaschke, Olaf. 1998. "Die Elimination Wissenschaftlicher Unterscheidungsfähigkeit: Goldhagens Begriff des 'Eliminatorischen Antisemitismus'—eine Überprüfung." In *Geschichtswissenschaft und Öffentlichkeit. Der Streit um Daniel J. Goldhagen*, ed. Johannes Heil and Rainer Erb. Frankfurt am Main: S. Fischer.

Braudel, Fernand. 1980. *On History*. Chicago: University of Chicago Press.

Browning, Christopher R. 1993. *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. New York: HarperCollins.

Capoccia, Giovanni and Daniel Ziblatt. 2010. "The Historical Turn in Democratization Studies." *Comparative Political Studies* 43(8–9): 931–68.

Carr, Edward Hallett. 1961. *What Is History?* New York: Vintage.

Carrier, Richard. 2012. *Proving History: Bayes's Theorem and the Quest for the Historical Jesus*. Amherst, NY: Prometheus Books.

Collier, David and Steven Levitsky. 1997. "Democracy with Adjectives: Conceptual Innovation in Comparative Research." *World Politics* 49(3): 430–51.

Coppedge, Michael, John Gerring, David Altman, Michael Bernhard, Steven Fish, Allen Hicken, and Matthew Kroenig. 2011. "Conceptualizing and Measuring Democracy: A New Approach." *Perspectives on Politics* 9(2): 247–67.

Daston, Lorraine. 1991. "Marvelous Facts and Miraculous Evidence in Early Modern Europe." *Critical Inquiry* 18(1): 93–124.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

———. 2001. "Why Facts Are Short." *Quadernistorici* 36(3): 745-70.
———. 2011. "The Empire of Observation, 1600-1800." In *Histories of Scientific Observation*, ed. Lorraine Daston and Elizabeth Lunbeck. Chicago: University of Chicago Press.
Daston, Lorraine and Elizabeth Lunbeck. 2011. *Histories of Scientific Observation*. Chicago: University of Chicago Press.
Duneier, Mitchell. 2011. "How Not to Lie with Ethnography." *Sociological Methodology* 41(1): 1–11.
Eley, Geoff. 2000. *The "Goldhagen Effect": History, Memory, Nazism—Facing the German Past*. Ann Arbor: University of Michigan Press.
Elman, Colin and Diana Kapiszewski. 2014. "Data Access and Research Transparency in the Qualitative Tradition." *PS: Political Science & Politics* 47(1): 43–47.
Evans, Richard. 1997. *In Defense of History*. New York: W. W. Norton.
Fischer, David Hackett. 1970. *Historians' Fallacies: Toward a Logic of Historical Thought*. New York: Harper & Row.
Fischhoff, Baruch. 1982. "For Those Condemned to Study the Past: Heuristics and Biases in Hindsight." In *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic and Amos Tversky. Cambridge: Cambridge University Press.
Fleck, Ludwik. 1935. *Genesis and Development of a Scientific Fact*. Chicago: University of Chicago Press.
Gaddis, John Lewis. 2002. *The Landscape of History*. Oxford: Oxford University Press.
Geertz, Clifford. 1973. "Thick Description: Toward an Interpretive Theory of Culture." In *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.
George, Alexander and Andrew Bennett. 2004. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.
Gerring, John. 2012a. *Social Science Methodology*. 2d ed. New York: Cambridge University Press.
———. 2012b. "Mere Description." *British Journal of Political Science* 42(4): 721–46.
Gerring, John and Dino Christenson. 2017. *Applied Social Science Methodology: An Introductory Guide*. Cambridge: Cambridge University Press.
Goldhagen, Daniel Jonah. 1992. "The Evil of Banality." *New Republic* 207(3/4): 49–52.
———. 1996a. *Hitler's Willing Executioners. Ordinary Germans and the Holocaust*. New York: Knopf.
———. 1996b. "Motives, Causes, and Alibis." *New Republic*, December 23, 37–45.
———. 1997. "The Fictions of Ruth Bettina Birn." *German Politics and Society* 15(3): 119–65.
Grafton, Anthony. 1999. *The Footnote: A Curious History*. Cambridge, MA: Harvard University Press.
Graves, Lucas. 2016. *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*. New York: Columbia University Press.
Haggard, Stephan and Robert R. Kaufman. 2012. "Inequality and Regime Change: Democratic Transitions and the Stability of Democratic Rule." *American Political Science Review* 106(3): 495–516.
Hagstrom, Robert. 2013. *Investing: The Last Liberal Art*. New York: Columbia University Press.
Hall, Peter. 2003. "Aligning Ontology and Methodology in Comparative Politics." In *Comparative Historical Analysis in the Social Sciences*, ed. James Mahoney and Dietrich Rueschemeyer. Cambridge: Cambridge University Press.
Hand, David. 2016. *Measurement: A Very Short Introduction*. New York: Oxford University Press.
Heil, Johannes and Rainer Erb, eds. 1998. *Nachgelesen. Goldhagen und Seine Quellen*. Frankfurt: S. Fischer.
Howard, Christopher. 2017. *Thinking Like a Political Scientist: A Practical Guide to Research Methods*. Chicago: University of Chicago Press.
Howell, Martha and Walter Prevenier. 2001. *From Reliable Sources: An Introduction to Historical Methods*. Ithaca, NY: Cornell University Press.
Hurst, B. C. 1981. "The Myth of Historical Evidence." *History and Theory* 20(3): 278.
King, Gary. 2013. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.
King, Gary, Robert Keohane and Verba Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
Knapp, Peter. 1984. "Can Social Theory Escape from History? Views of History in Social Science." *History and Theory* 23(1): 34–52.
Kocher, Matthew and Nuno Monteiro. 2016. "Lines of Demarcation: Causation, Design-Based Inference, and Historical Research." *Perspectives on Politics* 14(4): 952–74.
Kreuzer, Marcus. N.d. "Unwitting Bayesians. The Bayesian Foundation of Historical Analysis." Working paper.
Kreuzer, Marcus. 2003a. "Parliamentarization and the Question of German Exceptionalism, 1867–1918." *Central European History* 36(3): 327–59.
———. 2003b. "Reply to Sperber and Ledford." *Central European History* 36(3): 375–82.
———. 2010. "Historical Knowledge and Quantitative Analysis: The Case of the Origins of Proportional Representation." *American Political Science Review* 104(2): 369–92.
———. 2016. "Assessing Causal Inference Problems with Bayesian Process Tracing: The Economic Effects of Proportional Representation and the Problem of Endogeneity." *New Political Economy* 22(5): 473–83.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press

———. 2018. "Data for: The Structure of Description: Evaluating Historical Description and Its Role in Theorizing." *Qualitative Data Repository*: https://doi.org/10.5064/F6OAEIDA.
Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.
Ledford, Kenneth. 2003. "Comparing Comparisons: Disciplines and the Sonderweg." *Central European History* 36(3): 367–74.
Lieshout, Robert, Mathieu Segers and Johanna Vleuten. 2004. "De Gaulle, Moravcsik, and The Choice for Europe: Soft Sources, Weak Evidence." *Journal of Cold War Studies* 6(4): 89–139.
Lupia, Arthur and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47(1): 19–42.
Lustick, Ian. 1996a. "Read My Footnotes." *ASPA-CP Newsletter* 7(1): 6, 10.
———. 1996b. "History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias." *American Political Science Review* 90(3): 605–18.
Mahoney, James. 2004. "Comparative-Historical Methodology." *Annual Review of Sociology* 30(1): 81–101.
Mahoney, James and Michael Ellsberg. 1999. "Goldhagen's *Hitler's Willing Executioners*: A Clarification and Methodological Critique." *Journal of Historical Sociology* 12(4): 422–36.
Markoff, John. 2002. "Archival Methods." In *International Encyclopedia of the Social and Behavioral Sciences*, ed. Neil Smelser and Paul Balteseds. Oxford: Elsevier.
McCullagh, C. Behan. 2000. "Bias in Historical Description, Interpretation, and Explanation." *History and Theory* 39(1): 39–66.
Moravcsik, Andrew. 2013. "Did Power Politics Cause European Integration? Realist Theory Meets Qualitative Methods." *Security Studies* 22(4): 773–90.
Morgan, Mary S. 2011. "Seeking Parts, Looking for Wholes." In *Histories of Scientific Observation*, ed. Lorraine Daston and Elizabeth Lunbeck. Chicago: University of Chicago Press.
Moses, A. Dirk. 1998. "Structure and Agency in the Holocaust: Daniel J. Goldhagen and His Critics." *History and Theory* 37(2): 194–219.
Narang, Vipin and Rebecca M. Nelson. 2009. "Who Are These Belligerent Democratizers? Reassessing the Impact of Democratization on War." *International Organization* 63(2): 357–79.
Ogilvie, Brian W. 2008. *The Science of Describing: Natural History in Renaissance Europe*. Chicago: University of Chicago Press.
Parsons, Craig. 2013. "Power, Patterns, and Process in European Union History." *Security Studies* 22(4): 791–801.
Pierson, Paul. 2003. "Big, Slow-Moving and Invisible: Macrosocial Processes in the Study of Comparative Politics." In *Comparative Historical Analysis in the Social Sciences*, ed. James Mahoney and Dietrich Rueschemeyer. Cambridge: Cambridge University Press.
Pohl, Dieter. 1997. "Die Holocaust-Forschung Und Goldhagens Thesen." *Vierteljahrshefte Für Zeitgeschichte* 45(1): 1–48.
Poovey, Mary. 1998. *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. Chicago: University of Chicago Press.
Porter, Theodore M. 1986. *The Rise of Statistical Thinking, 1820–1900*. Princeton, NJ: Princeton University Press.
Ragin, Charles. 1994. *Constructing Social Research*. Thousand Oaks, CA: Pine Forge Press.
Rosato, Sebastian. 2003. "The Flawed Logic of Democratic Peace Theory." *American Political Science Review* 97(4): 585–602.
Rosenberg, Daniel and Anthony Grafton. 2010. *Cartographies of Time: A History of the Timeline*. New York: Princeton Architectural Press.
Rueschemeyer, Dietrich. 2009. *Usable Theory: Analytic Tools for Social and Political Research*. Princeton, NJ: Princeton University Press.
Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *The American Political Science Review* 64(4): 1033–53.
Sewell, William. 1996. "Three Temporalities: Toward an Eventful Sociology." In *The Historic Turn in the Human Sciences*, ed.Terrence J. McDonald. Ann Arbor: University of Michigan Press.
Sewell, William H. 1967. "Marc Bloch and the Logic of Comparative History." *History and Theory* 6(2): 208–18.
Shandley, Robert R., ed. 1998. *Unwilling Germans? The Goldhagen Debate*. Minneapolis: University of Minnesota Press.
Silver, Nate. 2012. *The Signal and the Noise*. New York: Penguin.
Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge: Cambridge University Press.
Sperber, Jonathan. 2003. "Comments on Marcus Kreuzer's Article." *Central European History* 36(3): 359–67.
Stern, Fritz. 1996. "The Goldhagen Controversy: One Nation, One People, One Theory?" *Foreign Affairs* 75(6): 128–38.
Swedberg, Richard. 2012. "Theorizing in Sociology and Social Science: Turning to the Context of Discovery." *Theory and Society* 41: 1–40.
———. 2014. *The Art of Social Theory*. Princenton, NJ: Princeton University Press.

Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.
Trachtenberg, Marc. 2006. *The Craft of International History: A Guide to Method*. Princeton, NJ: Princeton University Press.
Wittenberg, Jason, ed. 2013. "What Do We Mean by Historical Legacies?" *Qualitative and Multi-Method Research Newsletter* 11(2): 8–9.
Wootton, David. 2015. *The Invention of Science: A New History of the Scientific Revolution*. New York: Harper Collins.
Zaks, Sherry. 2017. "Relationships Among Rivals (RAR): A Framework for Analyzing Contending Hypotheses in Process Tracing." *Political Analysis* 25(3): 1–19.
Zerubavel, Eviatar. 2003. *Time Maps: Collective Memory and the Social Shape of the Past*. Chicago: University of Chicago Press.

https://doi.org/10.1017/S1537592718001068 Published online by Cambridge University Press