AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Qualitative Imputation of Missing Potential Outcomes

Alexander Coppock Yale University
Dipin Kaur Yale University

Abstract: *We propose a framework for meta-analysis of qualitative causal inferences. We integrate qualitative counterfactual inquiry with an approach from the quantitative causal inference literature called extreme value bounds. Qualitative counterfactual analysis uses the observed outcome and auxiliary information to infer what would have happened had the treatment been set to a different level. Imputing missing potential outcomes is hard and when it fails, we can fill them in under best- and worst-case scenarios. We apply our approach to 63 cases that could have experienced transitional truth commissions upon democratization, eight of which did. Prior to any analysis, the extreme value bounds around the average treatment effect on authoritarian resumption are 100 percentage points wide; imputation shrinks the width of these bounds to 51 points. We further demonstrate our method by aggregating specialists' beliefs about causal effects gathered through an expert survey, shrinking the width of the bounds to 44 points.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/2IVKXD.

A special promise of qualitative counterfactual inquiry is that deep case knowledge can generate informed guesses of *what would have happened had things been different*. We observe realized outcomes but by definition we can never observe counterfactual outcomes – they are counter-to-fact. Therefore, a causal inference in a single case amounts to a claim about the value of a missing potential outcome. We refer to such claims as 'imputations'.

We acknowledge from the outset that counterfactual analysis is not usually the only purpose of qualitative research, but it is often at least one of the purposes, and this article is about those purposes only. Our goal is to show how a systematic aggregation of counterfactual guesses can shrink fundamental uncertainty about average causal effects in a principled manner. Extreme value bounds (Manski 1999) represent the logical range of average causal effects that are consistent with the world as we observe it. The bounds start out quite wide, but we show in this article how qualitative imputation of missing potential outcomes can tighten the bounds substantially.

Our approach shares much in common with those of Seawright (2016), Glynn and Ichino (2015) and Humphreys and Jacobs (2015), each of whom incorporate qualitative and quantitative information to produce better *estimates* of average causal effects. By contrast, we leave entirely to the side the question of when and how qualitative researchers should draw causal inferences in a single case. Doing so is generally quite difficult, but obviously not impossible. Qualitative methodologists have developed a battery of approaches for single-case causal inference, and different cases may call out for different methods (see, e.g. Goertz and Mahoney 2012; George and Bennett 2005; Bennett and Checkel 2015; Ragin 2014). We seek to incorporate single-case causal inferences (no matter their specific methodological provenance) within a synthesized framework, something akin to a meta-analysis for qualitative inquiry.

Alexander Coppock is Assistant Professor of Political Science, Yale University (alex.coppock@yale.edu). Dipin Kaur is a PhD candidate in the Department of Political Science, Yale University (dipin.kaur@yale.edu).

The authors thank Peter Aronow, Graeme Blair, Angele Delavoye, Morgan Galloway, Michael Goldfein, Matthew Graham, Macartan Humphreys, Cleo O'Brien-Udry, Jason Seawright, Drew Stommes, Elisabeth Wood and Lauren Young for helpful conversations and discussions. This study received approval from the Yale University Institutional Review Board (IRB 2000029290) for the survey of expert opinions.

American Journal of Political Science, Vol. 00, No. 0, xxxx 2022, Pp. 1–15

DOI: 10.1111/ajps.12697

Much recent work has sought to combine quantitative and qualitative methodologies for causal inference (Brady and Collier 2010; Blair et al. 2019; Glynn and Ichino 2015; Humphreys and Jacobs 2015; King, Keohane, and Verba 1994; Mahoney 2010), often through a potential outcomes framework (Acharya, Blackwell, and Sen 2016; Abadie, Diamond, and Hainmueller 2015, 2010; Imai et al. 2011; Morgan and Winship 2014; Pearl 2009; Seawright 2016). However, counterfactual analysis has a long tradition among qualitative methodologists as well (Hume 1748; Lewis 1979, 1973; Woodward 2005). Three examples from recent empirical scholarship illustrate the diverse range of approaches to counterfactual analysis used by qualitative researchers.[1]

Harvey (2012) evaluates the conventional wisdom that the U.S. decision to invade Iraq in 2003 was a product of neoconservative ideology, internal delusions and grand strategies unique to President George W. Bush and his national security team. Here, the Bush presidency is the treatment condition, a counterfactual Al Gore presidency is the control condition and the outcome of interest is the Iraq War. The world revealed the treatment outcome: $Y(\text{Bush}) = \text{War}$. Harvey leverages a comparative counterfactual framework to impute the unobserved control outcome, $Y(\textit{Gore})$. To this end, he analyses 'facts and evidence derived from a careful (and complete) review of the relevant historical record' (Harvey 2012) including analyses of interviews, speeches and public statements. The author finds little support in favour of the generally accepted view that a Gore administration would have remained at peace. Instead, he concludes that the onset of the Iraq War was a product of many key decisions and entrenched misconceptions that constituted a path-dependent sequence of moves that pushed the U.S.–U.K. coalition to war, thereby implying that $Y(\text{Gore}) = \text{War}$ as well.

Haber and Menaldo (2011) operationalize explicitly specified counterfactuals to evaluate the resource curse hypothesis that countries with natural resource dependence are less likely to be democratic. The treatment here is reliance on natural resources and the outcome is regime type. To evaluate this relationship, Haber and Menaldo specify the counterfactual path that a resource-reliant country would have followed in the absence of those resources on the basis of the path followed by non–resource-reliant countries in its geographic region. They then compare the paths to see whether any divergence is correlated with differences in resource reliance. They conclude, contrary to conventional wisdom, that an increase in natural resource reliance does not promote authoritarianism. In other words, they claim that $Y(\text{Resources}) = Y(\text{No Resources})$.

Lastly, Lebow and Stein (1996) assess the counterfactual claim (asserted by, among others, Nikita Khrushchev) that had the Soviet Union not deployed missiles in Cuba triggering the Cuban Missile Crisis, the United States would have invaded the island. In this case, the treatment is the Soviet deployment of missiles and the outcome is American invasion into Cuba. Although history has revealed the treated outcome, Lebow and Stein use recently uncovered evidence to impute the unobserved untreated outcome. They show that even before the missile deployment, no influential members of the Kennedy administration wanted to attack Cuba. Kennedy and Secretary of Defense McNamara were impressed by the level of Cuban popular support for Fidel Castro and the defensive ability of the Cuban militia. Hence, they were deterred by the revised intelligence estimates, which indicated that a successful invasion would have required massive U.S. forces to remain in an occupational role for an indefinite period. Based on these costs, they resolved not to attack. Lebow and Stein conclude that Khrushchev was wrong: Soviet missiles were not necessary to prevent an American attack. If anything, the missile deployment only increased the political pressures on Kennedy to invade. In this case, $Y(\text{Missiles}) = \text{No invasion}$; Lebow and Stein (1996) claim $Y(\text{No missiles}) = \text{No invasion}$ as well.

The foregoing examples demonstrate that our approach is emphatically mixed-methods. Each of the three causal inferences relies on qualitative, within-case information to generate counterfactual claims. Examples 1 and 3 used process tracing whereas example 2 relied on case comparison. The beauty of describing causal effects in terms of counterfactuals is that we can express the research findings in a common language, regardless of the precise qualitative inference procedure used to generate the findings. Once the causal effects in a literature have been expressed in that common language, they can be straightforwardly synthesized.

Our approach leaves open the possibility that imputation may fail in a particular case. The researcher may conclude that the existing information is too thin or too heavily disputed to merit a confident counterfactual guess and may decide to leave a counterfactual outcome unimputed. For instance, despite the wealth of literature on the outbreak of the First World War, the question of which country started the war has been debated for over a century with no clear conclusions (Fischer 1967; Mombauer 2013). In such cases of deeply disputed causal

[1] See Fearon (1991) for examples of single-case counterfactual analysis related to the non-occurrence of World War III and regime types in Latin America and interwar Europe.

claims, counterfactual imputation may simply be out of reach.

In other cases, however, the analyst may express an *uncertain* guess (in the form of a probability statement) about counterfactual outcomes; the resulting meta-analysis incorporates that uncertainty. Our approach combines counterfactual imputation with extreme value bounds (explained in detail in the next section). The main purpose of the method is to structure and characterize the uncertainty surrounding average causal effects. The bounds reflect fundamental uncertainty because they are a function of what we know for sure (the data revealed by the world), what we think we know (the counterfactual inferences we draw in some cases) and what *we know* we do not know (the counterfactual outcomes in cases we leave unimputed).

Describing Counterfactuals Using Potential Outcomes

Under the Neyman–Rubin causal model (Neyman 1923; Rubin 1974), units are endowed with a set of potential outcomes, only one of which they reveal depending on the realization of exposure to causal factors. In the basic case, a unit i has only two potential outcomes, $Y_i(1)$ and $Y_i(0)$, which correspond to the treated and untreated potential outcomes. This setup embeds the 'stable unit treatment value assumption' or SUTVA, which requires that unit i not have potential outcomes beyond $Y_i(1)$ and $Y_i(0)$ that might depend on the treatment assignments of other units. The realized treatment d_i then 'reveals' the observed outcome Y_i via the 'switching' equation: $Y_i = Y_i(1)d_i + Y_i(0)(1 - d_i)$. If treated, unit i reveals $Y_i(1)$ and if untreated, it reveals $Y_i(0)$.

The fact that we can never observe a unit in both its treated and untreated states has been famously dubbed the 'fundamental problem of causal inference' (Holland 1986). In Table 1, we see a treated unit ($d_i = 1$) and its observed outcome ($Y_i = 1$). We know $Y_i(1)$ equals 1 because the revealed outcome Y_i equals 1. The untreated potential outcome $Y_i(0)$ is missing. The goal of counterfactual analysis is to fill in the missing value with a (hopefully very well-educated) guess. Because the outcome in this example is binary, imputation amounts to filling in the question mark with either a zero or a one.

TABLE 1 Causal Inference for a Single Unit

d_i	Y_i	$Y_i(0)$	$Y_i(1)$
1	1	?	1

Note: This table represents the Fundamental Problem of Causal Inference that we do not observe units under conditions that did not occur, in this case the untreated condition.

One of the analytic tasks of qualitative research is to understand the separate impacts of the many causal factors that explain outcomes in a single case. Moreover, qualitative researchers often consider the mechanisms by which treatments affect outcomes. Both questions—*which* factors matter and *why*—are important, complicated and difficult to answer. We will focus on a tiny slice of that analytic task: understanding the impact of a single factor on a single outcome in a single case, inclusive of any and all mechanisms that may be at play. We study one causal factor out of the many that may determine an outcome because understanding just one is hard enough. An exhaustive accounting of all of the causes of an outcome is unachievable in most cases. Secondly, we are after the total effect of the treatment on the outcome rather than the portions of the total effect that can be attributed to various intermediate processes. For a discussion of the extreme difficulty inherent in studying mechanisms (mediators), even when treatments are randomly allocated to subjects, see Bullock, Green, and Ha (2010) or Gerring (2010).

Causal inference for a single unit requires the researcher to impute the missing potential outcome. Using case knowledge, information about individual actors' incentives, institutional arrangements, temporal variation and logic, qualitative researchers can make a guess about what would have happened had the treatment been set to a different level. The uncertainty attending to that guess can also be qualitatively expressed. For some units, this task is easy. For others, it is much harder. Qualitative methodologists have developed a suite of approaches for determining whether the available qualitative data are sufficient to license a causal inference (Bennett and Checkel 2015; Beach and Pedersen 2016; Collier 2011; Fairfield 2013; George and Bennett 2005; Gerring 2006; Mahoney 2010; Goertz and Mahoney 2012; Mahoney 2012; Rohlfing 2012; Ragin 2014).

The most common approach is the 'counterfactual case strategy', which Fearon (1991) wryly notes, 'often goes under the name "case study"'. In the counterfactual case strategy, researchers imagine a counterfactual world in which the presumed causal factor is absent but everything else is identical. Researchers then support their causal claims by either *invoking other causal claims*—laws, regularities or principles that have independent credibility—or by drawing on knowledge of relevant historical facts. In our analysis of truth commissions, we most often follow the counterfactual case strategy. In the process, we pay close attention to the six attributes

TABLE 2 Causal Types with Binary Treatments and Outcomes

Type	$Y_i(0)$	$Y_i(1)$	τ_i
*A*dverse	1	0	-1
*B*eneficial	0	1	1
*C*hronic	0	0	0
*D*estined	1	1	0

Note: Each row describes a causal type in terms of its untreated potential outcome, its treated potential outcome, and the difference between them.

of robust counterfactual thought exercises developed in Tetlock and Belkin (1996).

Although the counterfactual case approach compares observed cases with imagined counterfactual cases, formalized process tracing examines diagnostic pieces of evidence within individual cases to develop and test hypotheses about causal mechanisms (Bennett and Checkel 2015). Process tracing involves the application of various empirical tests to adjudicate among alternative explanations. These tests focus on evidence with different kinds of probative value, and go by the names of hoop tests (necessary but not sufficient), straw in the wind tests (neither necessary nor sufficient), smoking gun tests (sufficient but not necessary) and doubly decisive tests (necessary and sufficient) (Bennett 2010; Van Evera 1997). Once all the processes that connect a treatment to an outcome have been traced, counterfactual scenarios can be imputed, yielding an estimate of the total treatment effect. Some of the authors we rely on to make our imputations (such as Wiebelhaus-Brahm 2020; Bakiner 2015) take process tracing approaches to the study of truth commissions in Chile, Peru, Sri Lanka, Morocco and Bahrain.

Recent developments in process tracing have incorporated a specifically Bayesian perspective to single-case analysis. For example, the Bayesian Integration of Quantitative and Qualitative (BIQQ) data framework developed by Humphreys and Jacobs (2015) imagines four causal types defined by the potential outcomes a unit would express depending on the presence or the absence of treatment.[2] These types are shown in Table 2: Chronic and destined types experience no effect of treatment because their treated and untreated potential outcomes are equal. Adverse types experience a negative effect of treatment and beneficial types experience a positive effect of treatment.

[2]See Fairfield and Charman (2017) for a critique of the BIQQ procedure on the grounds that prior probabilities are often difficult to specify in some applied settings.

In this framework, analysts use values of the observed treatment and outcome to narrow down the possible causal types to just two; the output of the BIQQ process is a posterior probability that a case is one of two possible types. For example, if we observe $d = 1$ and $Y_i = 1$, we know the unit must either be a beneficial or a destined type. We then proceed to find 'clues' or within-case evidence (gathered using qualitative process tracing and quantitative data) to update our prior beliefs about case types and associated probabilities. Suppose we put the prior probability of being a beneficial type at .6, the probability of seeing the clue if it is a beneficial type at .9 and the probability of seeing the clue if it is an adverse type at .2. We then empirically observe the clue and update our beliefs using Bayes' rule:

$$
\begin{aligned}
&P(type = B|clue) \\
&= \frac{P(clue|type = B) * P(type = B)}{P(clue|type = B) * P(type = B) + P(clue|type = A) * P(type = A)} \\
&= \frac{0.9 * 0.6}{0.9 * 0.6 + 0.2 * 0.4} \\
&= 0.87.
\end{aligned}
$$

In other words, our posterior probability that the missing potential outcome $Y_i(0)$ equals 0 is 0.87. As we show below, our framework can seamlessly incorporate probabilistic guesses about missing potential outcomes when calculating extreme value bounds around average causal effects.

In our empirical application, we rely in large part on (our reading of) the qualitative inferences drawn by other researchers, each of whom has made choices among the variety of methods available to them. As our application below shows, sometimes no approach (qualitative, quantitative or otherwise) is sufficient for causal inference and we are forced to admit ignorance of causal effects. We view the ability to incorporate the *lack* of knowledge about counterfactuals as a major strength of our procedure.

The Procedure

Our goal will be to summarize qualitative inferences about individual-level causal effects for a set of N units. In particular, we aim to place bounds around the average treatment effect (ATE) for these N units: $ATE \equiv \frac{\sum_i^N Y_i(1) - Y_i(0)}{N}$. The ATE is a common target of inference in quantitative research but less so in qualitative work (Goertz and Mahoney 2012). One might reasonably argue that a chief advantage of qualitative methods is that

they are addressed to inferential targets that are far more subtle than a simple average over possibly very heterogeneous cases. In this article, we bound two estimands, the ATE and the ATE on the treated, but our procedure can easily be applied to many other targets, including conditional ATEs, or any estimand that can be expressed as a function of the joint distribution of potential outcomes.

We first describe the procedure with binary outcomes. In the extensions to follow, we consider nonbinary outcomes and probabilistic beliefs about counterfactuals.

Extreme value bounds (Manski 1999) are the logical bounds around the ATE.[3] Consider a setting with a binary outcome and a binary treatment. In the 'best' case, the outcome for treated units is '1' and the outcome control units is '0'. In this scenario, the ATE is +100 percentage points. By the same logic, in the worst-case scenario, the ATE is −100 percentage points. Before any data are collected, the extreme value bounds are 200 percentage points wide.

Once empirical data are collected, we observe each unit in either its treated or untreated state. In the control group we observe $Y_i(0)$ and in the treatment group we observe $Y_i(1)$. If we now impute the best- and worst-case scenarios, we only have to impute *half* the potential outcomes because the world has revealed the other half. After data collection, the extreme value bounds shrink from 200 points wide to 100 points wide. These bounds represent—before the inclusion of any priors, qualitative information or other expertise—the uncertainty attending to the ATE. This uncertainty is not due to the sampling or assignment procedures, but instead due to our ignorance of the missing potential outcomes.

In order to further shrink the bounds, we impute missing potential outcomes using available qualitative case and process knowledge. In the empirical exercise below, we consider whole 'batches' of imputations at a time to structure the successive accumulation of evidence, but the order in which particular cases are imputed does not affect the final width of the bounds.

Extensions

Here we consider two extensions of the basic procedure.

First, some analysts may balk at the idea of giving a definitive counterfactual imputation, equivalent to asserting a counterfactual outcome with certainty. In such cases, we may be able to express a guess about counterfactuals in terms of a probability, as in the output of a BIQQ procedure (Humphreys and Jacobs 2015).

We can incorporate probabilistic beliefs with a slight elaboration of the bounding procedure. We divide unknown potential outcomes into three classes: those we can impute with certainty, those we cannot impute at all and those for which we can state the probability of the binary outcome taking on the value 1. If there are k potential outcomes in this third class, then we have to consider 2^k possible sets of potential outcomes. In each of the 2^k possibilities, we can compute extreme value bounds. The "point estimate" for the bounds is equal to its expectation, or the probability-weighted sum over all 2^k possibilities. We can then express our uncertainty as the 2.5th and 97.5 quantiles of the distribution of possibilities. In practice, this set may be enormous, in which case we can draw an arbitrarily large random sample from the full set of 2^k possibilities.

This extension highlights the difference between two sources of uncertainty: ignorance of the missing potential outcome and beliefs about the probability that the missing potential outcome equals 1. The first is about fundamental uncertainty. Because of the fundamental problem of causal inference, we do not know what would have happened had the treatment been set to a different level. The second is about a specific belief on the basis of case knowledge about the probabilities of counterfactual outcomes. To see this, suppose that an analyst claims that the probability of the missing potential outcome is .5. This 'coin flip' imputation is importantly different from making no imputation at all because the resulting extreme value bounds will be *tighter* than if the outcome were left unimputed altogether. Stated differently, claiming that the missing potential outcome has a probability of being a 1 of exactly 0.5 requires more qualitative information than leaving the outcome unimputed.

Second, we can extend our framework beyond binary outcomes to continuous or quasi-continuous variables. Suppose the logical extrema of the outcome variable are Y^{MAX} and Y^{MIN}. The width of the bounds can be computed as $\frac{[(Y^{MAX}-(Y^{MIN})*N_{unimputed}}{N}$, where $N_{unimputed}$ is the number of missing potential outcomes we decline to impute. We calculate the maximum and minimum possible values for the average treated ($\overline{Y_1}$) and untreated ($\overline{Y_0}$) potential outcomes as follows:

$$\overline{Y_1^{MAX}} = \frac{\sum_1^m Y_i + (N-m) * Y^{MAX}}{N},$$

$$\overline{Y_1^{MIN}} = \frac{\sum_1^m Y_i + (N-m) * Y^{MIN}}{N},$$

[3] Bounds are often used when outcome data are missing. See Gerber and Green (2012, chapter 7) for an accessible introduction to bounding approaches for attrition. Keele et al. (2017) show how to incorporate prior beliefs about the distribution of potential outcomes and unobserved confounding to shrink the extreme value bounds using a Bayesian approach.

$$\overline{Y_0^{MAX}} = \frac{\sum_{m+1}^{N} Y_i + (m) * Y^{MAX}}{N},$$
$$\overline{Y_0^{MIN}} = \frac{\sum_{m+1}^{N} Y_i + (m) * Y^{MIN}}{N},$$

where the first m of N units are treated and the remainder are untreated. The upper bound is $\overline{Y_1^{MAX}} - \overline{Y_0^{MIN}}$ and the lower bound is $\overline{Y_1^{MIN}} - \overline{Y_0^{MAX}}$.

Imputing continuous outcomes may be especially challenging, because the analyst is required to pick one value out of a range of values, rather than make a (comparatively) easier call about a binary state. A middle ground between imputing a particular value and leaving the outcome unimputed is to redefine what the best and worst cases are on a unit-by-unit basis. For example, imagine units are countries and the outcome is a 1–7 Freedom House score (the higher the score, the less free the country). Suppose the observed outcome for a treated country is 2, and the analyst's goal is to impute its untreated outcome. The analyst believes for good reasons that the treatment improved outcomes, but is unsure how much. For that unit only, we could define Y^{MAX} as 3 and Y^{MIN} as 2. The resulting extreme bounds around the ATE will be tighter than if we brought no information to bear at all.

Both extensions can be implemented simultaneously if the researcher can articulate a probability distribution for each missing potential outcome. For example, if we were confident that outcomes follow Poisson distributions, we could specify a particular λ (the parameter that governs the Poisson distribution) for each missing potential outcome. Computationally, we would estimate extreme value bounds many thousands of times, sampling from each missing potential outcome distribution each time. Taking the average of each bound will yield point estimates, and we can characterize our uncertainty with the 2.5th and 97.5th quantiles of the distribution of each bound.

A Toy Example

Consider a population of $N = 10$ units, seven of whom have been treated and three of whom have not. We observe the revealed (binary) outcome for all 10 units. The outcome for three of the treated units and one of the untreated units is 1; the outcome is equal to 0 for the remaining units. Before adding any qualitative information, the bounds on the ATE extend from -40 percentage points to 60 percentage points. Table 3 presents the table of potential outcomes as it proceeds through three rounds of imputation. Unknown and unimputed potential outcomes are represented with question marks and imputed outcomes are shown in bold red.

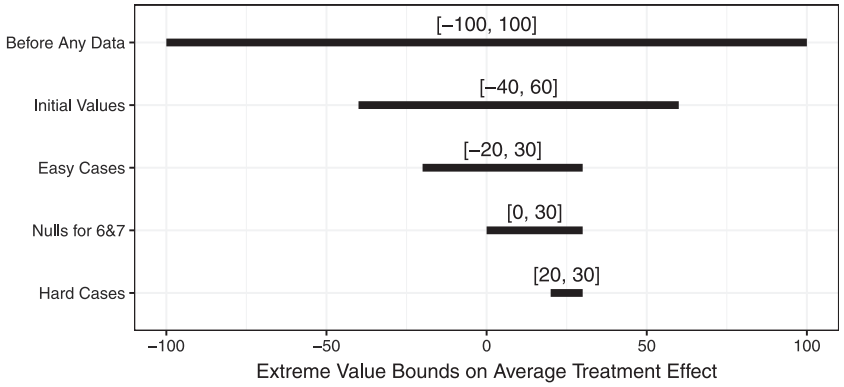
The fourth and fifth columns describe the imputation of five 'easy' cases. These are scenarios in which the untreated outcomes for units 1, 2, and 5 are obviously (to the researcher) 1, 1, and 0, respectively. Similarly, the

TABLE 3 Toy Example: Potential Outcomes Table

	Observed		Initial Values		Easy Cases		Nulls for 6 and 7		Hard Cases	
	d_i	Y_i	$Y_i(0)$	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$
1	1	1	?	1	**1**	1	1	1	1	1
2	1	1	?	1	**1**	1	1	1	1	1
3	1	1	?	1	?	1	?	1	**0**	1
4	1	1	?	1	?	1	?	1	**0**	1
5	1	0	?	0	**0**	0	0	0	0	0
6	1	0	?	0	?	0	**0**	0	0	0
7	1	0	?	0	?	0	**0**	0	0	0
8	0	1	1	?	1	**1**	1	1	1	1
9	0	0	0	?	0	**0**	0	0	0	0
10	0	0	0	?	0	?	0	?	0	?
	EV bounds:		[−40, 60]		[−20, 30]		[0, 30]		[20, 30]	

Note: Unknown and unimputed potential outcomes are represented with question marks. Imputations made on the basis of qualitative information are shown in bold red.

FIGURE 1 Toy Example: Extreme Value Bounds



Note: The extreme value bounds around the average treatment effect in this toy example shrink as unobserved potential outcomes are imputed on the basis of qualitative information.

treated outcomes of units 8 and 9 are 1 and 0. These cases are 'easy' in the sense that it is clear to the researcher that the treatment could not have had an effect on the outcome, perhaps because the outcome was clearly the consequence of a complex set of factors that exclude the treatment under consideration. These five imputations have shrunk the extreme value bounds considerably, and they now reach from −20 points to 30 points.

Suppose the researcher reasons next that the treatment should have, if anything, had a positive effect for units 6 and 7. Because the revealed treated outcomes were 0 for both units, the researcher concludes that the treatment must have had no effect on those units and imputes 0s for the untreated outcomes as well. This imputation shrinks the bounds to [0, 30].

As a final step, the researcher invests heavily in the remaining 'hard cases', units 3, 4 and 10. Suppose that the efforts are only partially rewarded. In cases 3 and 4, the researcher concludes that the treatment had a positive effect, and so imputes a 0 for the untreated outcomes in those cases. But in case 10, the empirical record is too thin and the causal story too murky to make a confident call. The resulting bounds [20, 30] correctly incorporate the remaining uncertainty in the researcher's summary of beliefs about causal effects.

Figure 1 shows the width of the extreme value bounds at each step in the process. The incorporation of empirical information and qualitative beliefs reduces the width of the bounds from 200 points to 10 points, corresponding to a dramatic reduction in fundamental uncertainty.

Probabilistic Extension

Suppose we are uncertain about the second step of the toy application. Instead of being sure that the untreated potential outcomes for units 6 and 7 are 0, we think the probabilities of being a '1' for units 6 and 7 are .3 and .4, respectively ($k = 2$). We now have to consider $2^2 = 4$ possibilities for units 6 and 7 which occur according to the researcher's probabilistic beliefs. Accordingly, there are four sets of extreme value bounds, as shown in Table 4. The point estimate for each bound is a probability-weighted average: [−5, 25]. We can characterize the uncertainty attending to the bounds with reference to the 2.5th and 97.5th quantiles of the distributions of each bound: [−20, 0] for the lower bound and [10, 30] for the upper bound. These bounds are (appropriately) much wider than the final bounds shown in Figure 1, because we have incorporated our newfound uncertainty about units 6 and 7.

TABLE 4 Application of Probabilistic Procedure to Toy Example

Unit 6	Unit 7	EV Bounds	Probability
0	0	[0, 30]	0.56
1	0	[−10, 20]	0.14
0	1	[−10, 20]	0.24
1	1	[−20, 10]	0.06

Note: Each row shows the extreme value bounds that would result depending on the missing potential outcomes for units 6 and 7, along with the researcher's belief about the probability of each scenario.

Application to the Average Effect of Truth Commissions Established at Democratization

In this section, we apply our procedure to the study of transitional truth commissions (TTCs), an area where a considerable amount of case-specific qualitative information has been generated by researchers within and without the academy. Studying the average effect of TTCs using standard quantitative methods like matching or regression with controls is probably too difficult because those places that come to be treated differ in so many ways (both observed and unobserved) from those places that remain untreated. In this setting, a claim that one has both measured and correctly adjusted for all possible confounders amounts to a leap of faith that is in our view not worth taking. We understand that opinions will differ on this point, but we want to note that it was our skepticism of the applicability of regression-like approaches that motivated the development of our procedure.

Our first task was to define the universe of cases that were *eligible* for a TTC. Clearly, all cases that received a TTC were eligible; the difficulty was finding those cases that could have but did not experience a TTC. We obtained information on transitions to democratic rule after a period of authoritarianism and arrive at a dataset of 63 observations that have a probability of experiencing a TTC between 0 and 1, exclusive.[4] We limited our search to the period from 1980 to 2008 because the first completed truth commission was established in 1983 and we need to allow some time to elapse after a truth commission is possible in order for the world to reveal outcomes. We identify democratic transition cases from the Geddes, Wright, and Frantz (2014) Autocratic Regimes Dataset. In this dataset, autocracy is defined as a set of formal or informal rules for choosing leaders and policies and each entry refers to consecutive years in which the same autocratic regime has been in power in a particular country.

Definition of Treatment and Outcomes

Our treatment is the establishment of a TTC. We use the definition of truth commissions given in Hayner (2000): 'officially created investigative bodies that document patterns of past human rights abuse over a specified period of time'. Of course, precisely which bodies meet this definition is disputed. Some lists of TTCs are relatively expansive (USIP 2011b; Olsen, Payne, and Reiter 2010; Hayner 2006) and others are relatively conservative (Dancy, Kim, and Wiebelhaus-Brahm 2010; Bakiner 2015; Kim and Sikkink 2010). In keeping with the more restrictive accounts, we consider a body a TTC if it (i) investigates for a limited amount of time, (ii) publishes a final report, (iii) examines a limited number of past events, (iv) enjoys autonomy from direct intervention by political actors, and (v) remains official in character (Bakiner 2015; Dancy, Kim, and Wiebelhaus-Brahm 2010).

Bakiner (2013) differentiates 'transitional' truth commissions (those that come up within the first three years of transition to peace or democracy) from 'non-transitional' truth commissions (those that do not arise in the context of transition), arguing that the two types display specific dynamics and require different analytic tools. In addition, Wiebelhaus-Brahm (2009b) describes non-transitional commissions as 'historic' truth commissions and argues that they are qualitatively different from commissions that take place during a transition. We consider as 'treated' the cases that Bakiner considers to be transitional.

The 'untreated' category includes not only cases that experience no commission whatsoever but also those whose commissions do not meet our criteria. For instance, countries that establish truth commissions many years after transition (Uruguay, South Korea, Panama, Brazil), those that announce truth commissions but do not implement them (Burundi) and those created by authoritarian governments (Morocco) are considered untreated. Also in this category are unofficial commissions created by community processes (such as Brazil), commissions of inquiry set up to investigate a singular event (such as a riot, pogrom or massacre leading to disappearances), and those set up to investigate embezzlement, fraud or similar crimes (as in Olsen et al. 2010). Finally, we consider disbanded commissions that are unable to complete their work (such as Bolivia or Ecuador) as untreated cases.

In randomized experiments, subjects are sometimes assigned to be treated but they fail to take treatment. Under an exclusion restriction that the assignment itself does not affect the outcome, these 'never-takers' reveal their *untreated* potential outcome. Accordingly, in cases that experienced commission-like bodies that do not meet our criteria, we assert (under an exclusion restriction) that they, like noncompliers, reveal their untreated potential outcomes. Our task then is to impute their genuinely *treated* potential outcomes. Our outcome of interest is the resumption of authoritarianism. We record outcomes as 1 when authoritarianism resumes within 10 years after transition, and 0 otherwise.

[4]See the Supporting Information (SI) for an application of our method to the study of truth commission in 54 post-conflict cases.

Imputing Missing Potential Outcomes

In this empirical example, we used the probabilistic version of the procedure. For all 63 unobserved potential outcomes, we imputed 5 with certainty, 26 probabilistically and we left 32 unimputed. We break up the imputations into four large categories. These categories represent the ease of imputation based on the depth of scholarship available in each case. We give the main reasons for our choices here and provide short descriptions of each case in the SI.

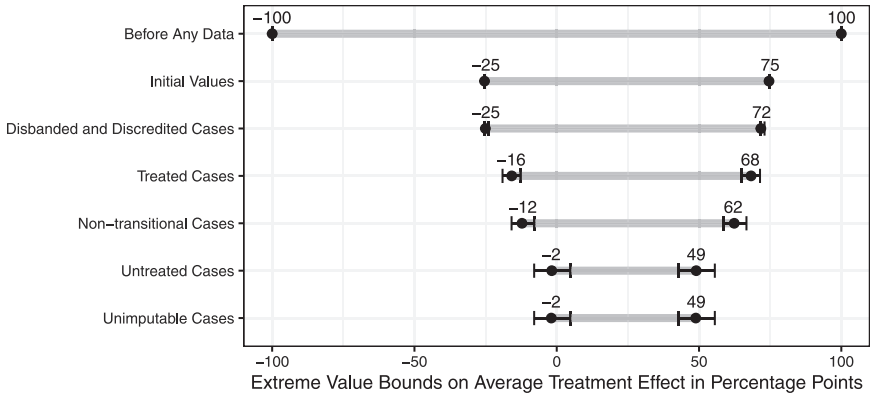
Step 1: Disbanded and Discredited Cases. We first identify truth commissions that were disbanded before completion (Bolivia in 1982 and the Philippines in 1986). These units reveal their *untreated* outcome. The observed outcome $Y_i(0)$ in each of these four cases was 0, as authoritarianism did not resume within 10 years of transition.

Reading the scholarship on each case, we found that they were disbanded due to underfunding or the lack of political will to investigate violations. We determined that they were not left incomplete or rendered ineffective because of fears of renewed authoritarianism. For instance, in the case of Bolivia, Hayner (2000), Skaar (1999) and USIP (2011a) find that the 1982 truth commission was unable to complete its work because of financial constraints. Before it was disbanded, however, the commission managed to document 155 cases of disappearance. Even though none of the cases were conclusively investigated, the attempt itself sparked numerous civil society debates. The ensuing public pressure eventually led political figures to set aside the initial amnesty law that protected the outgoing military regime from prosecution, leading to limited trials. We reason that because these alternative transitional justice mechanisms did not provoke an authoritarian backlash, it is unlikely that the truth commission, had it been carried to completion, would have catalysed a return to military dictatorship. In sum, had a truth commission been set up and completed in these cases, it would not have caused authoritarian resumption. We impute the treated potential outcome in each of these cases to be $Y_i(0) = Y_i(1) = 0$, with a probability of .1 that the imputed outcomes would take on the value 1 instead.

Step 2: Treated Cases. Next, we turn to the eight cases treated with bona fide TTCs. Most of these cases have been studied extensively, which allows us to draw on existing analyses from multiple sources. For instance, experts have cataloged data on the effects of the South African Truth and Reconciliation Commission on a variety of outcomes. Landmark studies by Gibson (2006, 2004, 2002) and Wiebelhaus-Brahm (2010) deem the truth commission as key to South Africa's transition. They find that dissatisfaction with the commission was largely limited to White South Africans, and more than 85% of Black South Africans interviewed believed that 'the commission did a reasonable job of letting families know what happened to their loved ones, of providing a true and unbiased account of the country's history, and of ensuring that human rights abuses would not happen again' (Gibson 2002). More importantly for our purposes, Wiebelhaus-Brahm (2010) conducted an explicit counterfactual analysis of the TTC's contribution to democracy, arguing: 'a brief counterfactual suggests that the TRC did play a significant role in this regard [contribution to democratic institutions]. Imagine a South Africa in which the TRC did not exist. Perhaps the [National Party] was able to extract a blanket amnesty as a concession for giving up power. Vigilantism would likely have exploded and whites would have fled South Africa in even larger numbers. Conversely, a South Africa in which many apartheid government officials were put on trial would seem a likely recipe for civil war. Many observers believed whites would prefer civil war to being ruled by the [African National Congress]. As it turned out, the TRC did just enough to satisfy all sides' (p. 48). In this context, even though the government's failure to institute timely and adequate reparations to victims immediately following the Truth Commission created renewed political tensions (Laplante and Theidon 2007), the investigation into crimes likely prevented their repetition. We therefore interpret the South African case as one where, in the absence of the TTC, the transition would have been incomplete and repression would have been very likely to resume. In other words, $Y_i(0) = 1$ with a probability of .8.

On the other end of the spectrum, we consider a case like Nigeria, which faced both moral and practical dilemmas in the setup of its Human Rights Violation Investigation Commission. First, the president who took over after the transition to democracy was himself once the leader of the military regime that was under investigation by the commission (Nwogu 2007; Yusuf 2007). In this and other similar cases of TTCs that completed their work but prioritized political ends over the updating of historical record and the creation of follow-up institutions, we think that the counterfactual outcome would be equal to the observed outcome. That is, we impute $Y_i(1) = Y_i(0) = 0$ with a probability of .1 that the imputed outcome would take on the value of 1.

FIGURE 2 Transitional Truth Commissions: Extreme Value Bounds



Note: The extreme value bounds around the average treatment effect shrink with the successive imputation of missing potential outcomes on the basis of qualitative information. We are uncertain of some imputations, so we use the probabalistic extension. The 95% confidence intervals characterize this additional source of uncertainty.

Step 3: Non-TTC Cases. We consider units that experience *non-transitional* commissions as untreated. That said, case study scholarship focusing on these non-transitional cases often provides clues about dynamics at the time of transition, allowing us to make guesses about what would have happened. In most scenarios where non-TTCs are set up many years later, we find that the demand for truth commissions existed even at the time of transition, but the reason for not establishing a truth commission was rarely a threat of renewed violence. Instead, the reason was often continued political infighting despite a formal transition, limited capacity amid other rebuilding concerns, leadership preferences or fatigue (USIP 2011b; Vandeginste 2012; Wiebelhaus-Brahm 2009a). Such cases are imputed as $Y_i(0) = Y_i(1)$. Under these conditions, a truth commission would neither have made things worse nor made them better.

Step 4: Untreated Cases. Lastly, we considered cases that were truly untreated. This step represents the toughest case for imputation, given the lack of scholarship about truth commissions that never occurred. Accordingly, most of these cases are left unimputed. The idea of truth commissions is sometimes brought up by civil society actors, opposition parties and international organizations but not acted upon. Some communities set up their own, unofficial truth-telling processes, whereas in others, we could find no evidence of discussion around transitional justice at all. Few studies actively address the reasons behind the failure of a truth commission to be set up, making our task more difficult.

Summary. Figure 2 summarizes our results. Before any data collection, the extreme value bounds are 200 points wide. After the world reveals half the potential outcomes, the width of the bounds shrink to 100 points. The four steps above shrink the uncertainty further as missing potential outcomes are filled in. The final bounds around the ATE are $[-2, 49]$ (51 points wide). The 95% confidence interval around each of these estimates expresses our uncertainty about the exact location of the bounds, according to our probabilistic evaluation of these counterfactual outcomes.

The bounds include zero. The data and our state of knowledge are currently consistent with positive, negative and zero average effects. This is *very importantly* different from a 'null' finding. The bounds are as wide as they are because we do not know as much about the effects of TTCs as we would like. The width of the bounds indicates either what work is left to be done or which counterfactuals are simply too unknowable to be imputed.

Table 5 reports the number of cases in which we think a TTC had or would have had a positive, negative or zero effect, as well as the number of cases in which we were unable to make an imputation. In our view, the most important pattern is that we impute non-zero causal effects only six times in these democratization cases. Another important pattern is that we are unable to make imputations in about half the total number of cases, hence the bounds remain wide and the gaps in our knowledge persist.

One of principal difficulties facing a traditional quantitative analysis of the effects of TTCs is that treated and untreated units differ from one another in both observed and unobserved ways. Accordingly, the ATE itself might not be the only interesting estimand—we might

TABLE 5 Summary of Imputations

Prevents Authoritarianism	No Effect	Causes Authoritarianism	Unimputed
2% (1)	40% (25)	8% (5)	51% (32)

Note: This table shows the percentage and number of our 63 cases falling into each of four imputation catagories.

wish to know the average effect of treatment among the treated (the ATT) or the average effect of treatment among the untreated (the ATU, sometimes called the ATC where the C stands for 'control'). The ATT and the ATU of course need not be equal.

Figure 3 shows how the extreme value bounds for the treated and untreated units develop differently. There are far more untreated units than treated units, and we know far less about them. The bounds on the ATU are greater than 58 points wide, compared with the bounds on the ATT, which shrink all the way down to a point. We can summarize the ATT as a −15 percentage point effect on return to authoritarianism.

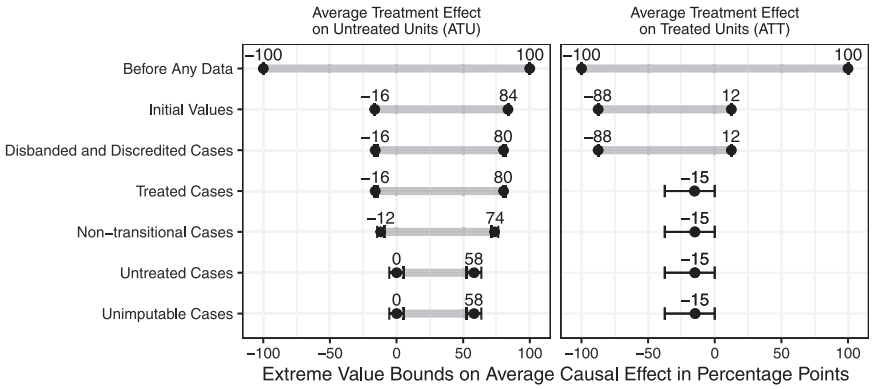
Expert Survey

As we have emphasized, scholars may reasonably disagree with our imputations. To demonstrate the advantages of our approach for surfacing and resolving such disputes, we conducted an email survey with country experts for each of our 63 cases. Identifying country experts is of course more art than science—our procedure was to search Google Scholar for 'transitional justice' and 'country X' to find qualified researchers. We received 20 responses to our survey, the text of which is reproduced in the SI.

We first assessed whether the experts agreed with our characterization of the observed outcome and the observed treatment status. All but five did agree. Those that did not either disputed that the country had ever truly transitioned to a functioning democracy or objected that countries we coded as untreated did experience a truth commission, albeit an incomplete one. We then asked experts to imagine what value the outcome variable would have taken had the treatment been set to the opposite level. The question had an explicit 'can't say' response option. The experts who objected to our coding of the untreated cases indicated that they explicitly answered remaining questions imagining that counterfactually, the country had been treated according to our definition. This generous survey-taking behaviour accords with our 'noncompliance' logic above.

Of these 20 cases, the imputations fully agreed with ours seven times (we either made the same imputations or both declined to impute) and directly conflicted three times (we made different imputations). In three cases, we imputed but the experts did not, and in the final seven cases, the experts imputed where we did not. Perhaps reflecting the experts' greater case knowledge, they gave imputations in 14 total cases compared with our 10. We view this level of agreement about a fundamentally unknowable quantity to be a qualified success.

FIGURE 3 Transitional Truth Commissions: Extreme Value Bounds by Treatment Status



Note: Cases are separated into those that experienced a truth commission (right panel) and those that did not (left panel).

Figure 4 shows how the bounds around the ATE reflect the expert imputations. We turn first to the set of 20 cases for which we have expert responses. Our original bounds for this set were 50 points wide; the bounds for the experts are 30 points wide, again reflecting their greater case knowledge. Among the full set of 63 cases, our original bounds were 51 points wide, but the experts' are 78 points wide. The extreme uncertainty reflects the 43 missing expert responses, which we leave as unimputed here. The last set of bounds in the figure reflects a combination of the expert evaluations in the 20 cases for which we received responses with our own evaluations in the remaining 43. The combined bounds are narrower than our original, reflecting a small but meaningful increase in cumulative beliefs about causal effects.

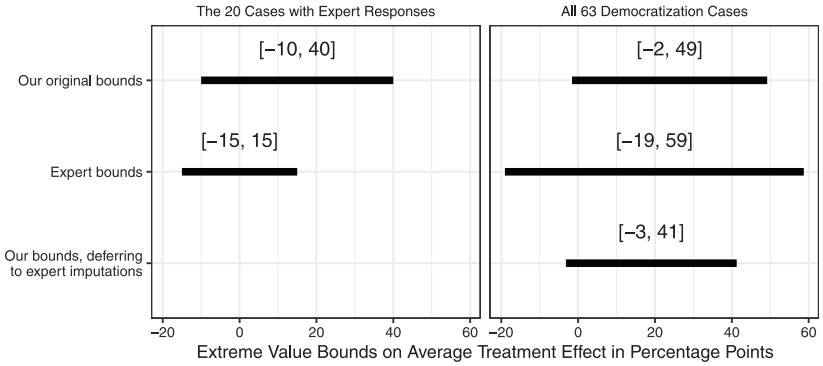
Surprisingly, all 20 of our respondents either made no imputation or imputed a counterfactual outcome that was equal to the observed outcome: all responses were either 'I don't know' or 'no effect'. One possible explanation is genuine skepticism that truth commissions affect regime type. This skepticism was expressed even in the case of South Africa, often considered the model for truth commissions worldwide. Although we deemed the South African truth commission as key to its transition based on qualitative and counterfactual analyses by multiple authors, our country expert claimed no effect, adding that 'my conclusion is based on the assumption that there was no transitional justice in the counterfactual scenario, and is largely based on the reality that the 1994 elections were peaceful, despite widespread predictions to the contrary, when there was no plan in place on what transitional justice would look like'. A second possibility is that the experts believe TTCs do have effects, but on outcomes beyond the one we considered.

When we asked our experts to explain their counterfactual reasoning, some respondents claimed that TTCs are 'endogenous to' stable regime types, in that they are established precisely when they will not endanger democracy. One respondent remarked that 'transitional justice mechanisms are the consequence, rather than the cause, of democratization processes' and another that 'truth commissions do not affect institutional democracy, because they are endogenous to ruling institutions'. These claims go beyond a worry about selection bias—they embed a causal belief that for truth commissions, treatment effects on the treated are zero because only countries with a zero treatment effect select into commissions.

We found an interesting asymmetry in our survey responses. Six experts declined to impute counterfactual outcomes, all six of which were untreated cases. By contrast, no expert assigned to a treated case declined to impute. Perhaps this pattern can be explained as an 'availability bias' in causal reasoning. It is apparently easier to imagine the absence of a treatment that did occur than to imagine the presence of a treatment that did not. One respondent who declined to impute expressed frustration at the difficulty of imagining when and how a commission that did not happen would have developed, writing, 'The levels of counter-factuality are quite mind-blowing'.

Indeed, our correspondence with experts prompted further reflection on the difficulties inherent in imagining the form of counterfactual conditions. In an untreated case, we have to imagine the form the TTC

FIGURE 4 Extreme Value Bounds Implied by Expert Imputations



Note: For the 20 cases in which experts made imputations (left panel), the extreme value bounds are narrower than our original bounds, reflecting the experts' deeper case knowledge. Combining expert beliefs about causal effects with our imputations (right panel) shrinks the width of the extreme value bounds around the average treatment effect for the full sample further to 44 points wide.

would take—and it is possible that the only plausible form would be so unlike the TTCs that did occur so as to render the comparison uninformative. We would exclude such cases under the rule that the universe of cases has to have a probability between 0 and 1 of experiencing a form of treatment that accords with our definition of TTCs. We want to limit the meta-analysis to cases for which we can imagine the relevant counterfactual conditions, even if we cannot impute counterfactual outcomes with confidence. As a side note, this conceptual difficulty applies to any mode of causal inference, qualitative or quantitative, but may go unnoticed when standard estimators are applied to multi-case datasets.

Discussion

In this article, we have proposed a procedure that combines single-case qualitative inference with extreme value bounds. The main purpose of the procedure is to meta-analyse qualitatively derived beliefs about average causal effects in a structured fashion. In cases where existing evidence is strong enough, we can impute counterfactual outcomes, which is equivalent to claiming knowledge of the individual treatment effect. In cases where existing evidence is weak, we consider worst- and best-case scenarios in order to places bounds on the ATE.

We think this procedure will be most applicable to medium-*N* empirical questions about which quite a bit of previous knowledge has been generated. This medium-*N* Goldilocks zone includes many topic areas in political science, including the study of leaders, states, or nations, intra- or inter-state armed conflicts, treaties or elections, to brainstorm a few. When the number of units is quite small, a series of case studies is probably more appropriate than our method. In larger-*N* settings, explicit counterfactual imputation for hundreds or thousands of missing potential outcomes on the basis of qualitative information may be challenging. One approach might be to sample from such cases, either at random or purposively, then zooming in to make imputations in a manageable set (as in the spirit of Lieberman 2005). A second approach might be to crowdsource the many imputations, as we demonstrated with our expert survey.

We faced some rewarding difficulties applying our method to TTCs. One trouble was defining potential outcomes in the first place. For those units that experienced a TTC, it is clear what is meant by the 'treated potential outcome'. But what does it mean if those places did not experience a TTC? Would they have experienced lustrations or purges instead? As alluded to above, it was more challenging to imagine the presence of a TTC in places where one did not occur. This asymmetry in our ability to imagine counterfactuals is troublesome, because decision makers contemplating a policy have to imagine the outcomes under many counterfactual scenarios that also have not occurred.

Of course, the main difficulty was making guesses about counterfactual states of the world. We were only able to engage with such a large set of cases because of the efforts of previous qualitative scholars of transitional justice. Our guesses about counterfactuals are summaries of our understanding of their work. We have laid out our reasoning for each case in the SI, but we are quite sure that others would dispute at least some of our guesses. Indeed, country experts disagreed with our imputations in three of 20 cases; ex post, we agree with them now, because we benefited from their expertise and reasoning.

Stepping back from the application of transitional justice, we think there are many advantages to summarizing qualitative inferences in this way. First, we avoid conditioning our analyses on treated cases only. Because units are not randomly assigned to treatments, the average effect of treatment among the treated need not be the same as the average effect of treatment among the untreated. Our approach avoids the distortions associated with studying treated units only.

Second, we can explicitly account for the non-random selection into treatment. If units that do and do not receive treatment are different from each other in both observed and unobserved ways, comparing them (as in an observational, quantitative study) is inappropriate. A series of single-case qualitative studies considers each unit individually, so worries about confounding are handled directly.

Third, the process is transparent. If critics disagree about an imputation, they can offer a different one. If the disagreement is insoluble, we can simply remove the imputation altogether. Disputes over imputations underscore that we do not know the counterfactual, so it would be inappropriate to claim knowledge about a particular causal effect. In the worst case, we would have to leave all counterfactuals unimputed, which could only occur if qualitative case knowledge were entirely useless for causal inference. We are skeptical that we can impute counterfactuals in all cases, but we are optimistic that we can do so in at least some.

Finally, we think that this procedure can serve as a way of making disagreements among scholars explicit. Oftentimes, alternative readings of a case have so little in common that even locating the source of disagreement is difficult. Using this procedure requires that scholars state their best guess as to what would have happened to an outcome in particular. In this way, it encourages scholars to be bold in their causal claims.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490): 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2): 495–510.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3): 512–29.

Bakiner, Onur. 2013. "Truth Commission Impact: An Assessment of How Commissions Influence Politics and Society." *International Journal of Transitional Justice* 8(1): 6–30.

Bakiner, Onur. 2015. *Truth Commissions: Memory, Power, and Legitimacy*. Philadelphia, PA: University of Pennsylvania Press.

Beach, Derek, and Rasmus Brun Pedersen. 2016. *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching, and Tracing*. Ann Arbor, MI: University of Michigan Press.

Bennett, A. 2010. "Process Tracing and Causal Inference." HE Brady and D Collier eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers: 207–220.

Bennett, Andrew, and Jeffrey T. Checkel. 2015. "Process Tracing: From Philosophical Roots to Best Practices." A Bennett and JT Checkel eds. *Process tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press: 3–37.

Bennett, Andrew, and Jeffrey T. Checkel, eds. 2015. *Process Tracing*. Cambridge: Cambridge University Press.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113(3): 838–59.

Brady, Henry E., and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. "Yes, But What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98(4): 550.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44(4): 823–30.

Dancy, Geoff, Hunjoon Kim, and Eric Wiebelhaus-Brahm. 2010. "The Turn to Truth: Trends in Truth Commission Experimentation." *Journal of Human Rights* 9(1): 45–64.

Fairfield, Tasha. 2013. "Going Where the Money Is: Strategies for Taxing Economic Elites in Unequal Democracies." *World Development* 47:42–57.

Fairfield, Tasha, and Andrew E. Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25(3): 363–80.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science - Jstor." *World Politics* 43(2): 169–95.

Fischer, Fritz. 1967. *Germany's Aims in the First World War*. New York: W.W. Norton.

Geddes, Barbara, Joseph Wright, and Erica Frantz. 2014. "Autocratic Breakdown and Regime Transitions: A New Data Set." *Perspectives on Politics* 12(2): 313–31.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gerring, John. 2006. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Gerring, John. 2010. "Causal Mechanisms: Yes, But?" *Comparative Political Studies* 43(11): 1499–1526.

Gibson, James L. 2002. "Truth, Justice, and Reconciliation: Judging the Fairness of Amnesty in South Africa." *American Journal of Political Science* 46(3): 540–56.

Gibson, James L. 2004. "Does Truth Lead to Reconciliation? Testing the Causal Assumptions of the South African Truth and Reconciliation Process." *American Journal of Political Science* 48(2): 201–17.

Gibson, James L. 2006. "Overcoming Apartheid: Can Truth Reconcile a Divided Nation?" *Annals of the American Academy of Political and Social Science* 603(1): 82–110.

Glynn, Adam N., and Nahomi Ichino. 2015. "Using Qualitative Information to Improve Causal Inference." *American Journal of Political Science* 59(4): 1055–71.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, NJ: Princeton University Press.

Haber, Stephen, and Victor Menaldo. 2011. "Do Natural Resources Fuel Authoritarianism? A Reappraisal of the Resource Curse." *American Political Science Review* 105(1): 1–26.

Harvey, Frank P. 2012. "President Al Gore and the 2003 Iraq War: A Counterfactual Test of Conventional 'W'isdom." *Canadian Journal of Political Science/Revue canadienne de science politique* 45(1): 1–32.

Hayner, Priscilla B. 2000. *Unspeakable Truths: Confronting State Terror and Atrocity*. New York: Routledge.

Hayner, Priscilla B. 2006. "Truth Commissions: A Schematic Overview." *International Review of the Red Cross* 88(862): 295–310.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60.

Hume, David. 1748. *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.

Humphreys, Macartan, and Alan M. Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109(4): 653–73.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4): 765–89.

Keele, Luke, Kevin M Quinn et al. 2017. "Bayesian Sensitivity Analysis for Causal Effects from 2X2 Tables in the Presence of Unmeasured Confounding with Application to Presidential Campaign Visits." *Annals of Applied Statistics* 11(4): 1974–97.

Kim, Hunjoon, and Kathryn Sikkink. 2010. "Explaining the Deterrence Effect of Human Rights Prosecutions for Transitional Countries." *International Studies Quarterly* 54(4): 939–63.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Laplante, Lisa J., and Kimberly Theidon. 2007. "Truth with Consequences: Justice and Reparations in Post-Truth Commission Peru." *Human Rights Quarterly* 29: 228.

Lebow, Richard Ned, and Janice Gross Stein. 1996. "Back to the Past: Counterfactuals and the Cuban Missile Crisis." P Tetlock and A Belkin, eds. Counterfactual Thought Experiments in World Politics. Princeton University Press: 119–48.

Lewis, David. 1973. "Counterfactuals and Comparative Possibility." W Harper, R Stalknaker and G Pearce eds. *Ifs*. Dordrecht, Holland: D. Reidel Publishing Company. pp. 57–85.

Lewis, David. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13(4): 455–476.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99(3): 435–52.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62(1): 120–47.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods and Research* 41(4): 570–97.

Manski, Charles F. 1999. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

et al Mombauer, Annika. 2013. "The Fischer Controversy 50 Years on." *Journal of Contemporary History* 48(2): 231–417.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Reprint, 1990." *Statistical Science* 5(4): 465–72. with Dabrowska, Dorota M. and Speed, Terence P.

Nwogu, Nneoma V. 2007. *Shaping Truth, Reshaping Justice: Sectarian Politics and the Nigerian Truth Commission*. Washington, DC: Lexington Books.

Olsen, Tricia D., Leigh A Payne, and Andrew G. Reiter. 2010. "Transitional Justice in the World, 1970-2007: Insights from a New Dataset." *Journal of Peace Research* 47(6): 803–09.

Olsen, Tricia D., Leigh A. Payne, Andrew G. Reiter, and Eric Wiebelhaus-Brahm. 2010. "When Truth Commissions Improve Human Rights." *International Journal of Transitional Justice* 4(3): 457–76.

Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.

Ragin, Charles C. 2014. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Los Angeles, CA: Univ of California Press.

Rohlfing, Ingo. 2012. *Case Studies and Causal Inference: An Integrative Framework*. London, UK: Palgrave Macmillan.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 688.

Seawright, Jason. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Skaar, Elin. 1999. "Truth Commissions, Trials-or Nothing? Policy Options in Democratic Transitions." *Third World Quarterly* 20(6): 1109–28.

Tetlock, Philip E., and Aaron Belkin. 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press.

USIP. 2011a. "Truth Commission: Bolivia - United States Institute of Peace." https://www.usip.org/publications/1982/10/truth-commission-bolivia

USIP. 2011b. "Truth Commission Digital Collection - United States Institute of Peace." https://www.usip.org/publications/2011/03/truth-commission-digital-collection

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Vandeginste, Stef. 2012. "Burundi's Truth and Reconciliation Commission: How to Shed Light on the Past While Standing in the Dark Shadow of Politics?" *International Journal of Transitional Justice* 6(2): 355–65.

Wiebelhaus-Brahm, Eric. 2009a. "What Does Brazil Have to Gain from a Truth Commission after Two Decades of Democracy?" *International Conference on the Right to Truth*, Sao Paulo, Brazil.

Wiebelhaus-Brahm, Eric. 2009b. "What Is a Truth Commission and Why Does It Matter?" *Peace and Conflict Review* 3(2): 1–14.

Wiebelhaus-Brahm, Eric. 2010. *Truth Commissions and Transitional Societies: The Impact on Human Rights and Democracy*. New York: Routledge.

Wiebelhaus-Brahm, Eric. 2020. "Global Transitional Justice Norms and the Framing of Truth Commissions in the Absence of Transition." *Negotiation and Conflict Management Research* 14(3): 170–186.

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Yusuf, Hakeem O. 2007. "Travails of Truth: Achieving Justice for Victims of Impunity in Nigeria." *The International Journal of Transitional Justice* 1(2): 268–86.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Application to End-of-Conflict Cases
Appendix B: Full Dataset
Appendix C: Democratization Imputations
Appendix D: Expert Survey