

Synthesizing Research through Total Model Space: The Case of Democracy and Growth

Andrés Cruz ^{*} John Gerring ^{*} Carl Henrik Knutsen [†]

April 13, 2024

Preliminary. Please do not cite or circulate. Comments welcome! [‡]

Abstract

Combining features associated with conventional literature reviews, meta-analyses, model-averaging, and robustness tests, we propose to learn from the assemblage of models contained in the literature and use them to construct an even more expansive model space. This total model space (TMS) is supposed to encompass all plausible tests of a specific research question. Using the example of democracy's suspected impact on growth, we show how the TMS can provide a more complete understanding of whether X is related to Y and how robust the relationship is. Moreover, our approach enables researchers to study which features contribute to a more optimistic or pessimistic conclusion, what combination of design choices are needed in order to reach standard thresholds of statistical significance, and how extant studies can be evaluated relative to that much larger set of plausible potential studies for the same research question. In this fashion, the TMS approach promises to contribute to the cumulation of knowledge, enabling researchers to narrow the range of potential disagreements and focus discussions on design choices that are most consequential.

^{*}University of Texas at Austin

[†]University of Oslo

[‡]We are grateful to Adam Glynn and Connor Jerzak as well as participants at the 2023 Annual APSA Meeting in Los Angeles for helpful comments and suggestions.

Introduction

Individual studies can be misleading. This is evident in the difficulties often found in replicating published work and in the diversity of findings that multiple published studies of the same subject often reveal ([Ioannidis, 2005](#)). Accordingly, one is well-advised to canvas the literature on a subject rather than to rely on individual studies before drawing conclusions. Trust literatures not papers, goes the adage.

What, then, is the best way to learn from the literature on a topic? How can numerous and often widely diverging findings be synthesized in a parsimonious and useful fashion? How can knowledge cumulation be facilitated?

The approach taken to this essential task depends, in part, on the sort of data upon which the literature is based. Here, we focus on studies using observational evidence with countries as the unit of analysis. Although crossnational data offers an especially hard case for causal inference ([Kittel, 2006](#); [Rodrik, 2012](#); [Seawright, 2007](#)), it is indispensable in addressing many of the (social science) questions that are of the greatest theoretical and practical interest. Growth, democracy, state capacity, inequality, war, and many other inputs and outputs make sense only (or at least primarily) at national levels. The theoretical and substantive importance of national units, coupled with their methodological difficulties, make this a critical setting for methodological reflection.

As a case in point, we focus on the relationship between democracy and growth. This question has attracted a great deal of scholarly attention over the years – by economists, political scientists and others – and exemplifies many of the characteristic features of crossnational data analysis, including diverse modeling choices and divergent findings ([Gerring, Knutsen, and Berge, 2022](#)).

To achieve cumulation, scholars usually rely on a literature review or meta-analysis. However, these approaches face several well-known obstacles. Published studies are prone to *p*-hacking and publication bias ([Brodeur, Cook, and Heyes, 2020](#)), and pre-analysis plans offer no deliverance because the relevant data is generally already in the public domain and hence cannot be credibly pre-registered. Accordingly, published studies offer what is, in all likelihood, a biased sample of all the analyses that have been conducted, or might be conducted. Complicating matters, crossnational studies often employ bespoke methodologies. Variations in the measurement of key concepts, background covariates, samples, and estimators make it difficult to compare results across studies. Finally, because samples overlap they are not independent and cannot be profitably combined (at least not without additional fixes)

in a standard meta-analysis. Given these caveats, it is not entirely clear how much one learns from a survey of published findings on questions such as democracy and growth.

A different approach to cumulation assembles the data and modeling choices that have accrued in the literature into a series of models that, collectively, are intended to represent all plausible ways of testing a causal question. These are then synthesized through some form of model-averaging (Montgomery and Nyhan, 2010; Steel, 2020). While this approach solves many problems associated with the traditional literature review or meta-analysis, it is not suitable for all purposes. First, model-averaging usually encompasses *all* the factors that might affect a given outcome — a causes-of-effects approach to causation. For outcomes with many contributing factors, this is immensely ambitious. Durlauf, Johnson, and Temple (2005) identifies 140 potential causes of economic growth, for example. As such, model-averaging is not well-designed for addressing the narrower question of X ’s impact on Y (an effects-of-causes approach to causation). Note that some background variables and estimators may be irrelevant for studying the effect of a particular X , and some specifications may introduce post-treatment bias. Second, model-averaging usually focuses on specification, leaving behind equally important choices in estimators, samples, and so forth. Finally, the all-important weighting function (how much weight to assign to each model) is usually determined by how the data support each model (using various measures of model-fit such as AIC, BIC, or RIC), measured in a Bayesian framework by posterior probabilities (perhaps with a few assumptions about g -priors). This approach is reasonable if the overall goal is predictive. If the goal is causal inference, however, it may lead one astray.

A third approach relies on robustness (or sensitivity) tests centered on a “benchmark” model, which is subject to myriad alterations – of specification, estimation, sample, data-processing, and so forth (Neumayer and Plümper, 2017; Young and Holsteen, 2017; Simonsohn, Simmons, and Nelson, 2020; Steegen et al., 2016). This protocol may not facilitate cumulation across a field for the simple reason that authors have different ideas about what that benchmark model should be, and there is no obvious way to resolve such differences. One hundred studies with one hundred different benchmark models will lead to one hundred sets of robustness tests, which are not directly comparable to each other since they enlist different samples, different measures of key concepts, different scope-conditions, and so forth.

Our initiative incorporates features from each of these traditions – literature re-

view, meta-analysis, model-averaging, and robustness tests. The core idea is to learn from the assemblage of models contained in a literature about an even more expansive model space that encompasses all plausible tests of a specific research question. By running all of these tests (regardless of whether they have been run previously by existing studies), and carefully observing and systematizing patterns in the results, we hope to gain a more complete picture of how X is related to Y – in this instance, how democracy is related to growth.

This approach, which we refer to as the *total model space* (TMS), is not intended to solve all knotty problems of causal inference encountered with observational data. However, it will help us to understand the literature on a subject, revealing which concatenations of modeling choices have been employed and which have been largely ignored, and allowing us to systematically compare results across these subsets. It will also reveal precisely which modeling assumptions are necessary for which set of conclusions, e.g., that democracy causes higher growth. In this fashion, TMS may contribute to the cumulation of knowledge, narrowing the range of potential disagreements and focusing discussion on design choices that are most consequential.

The article is organized according to the envisioned protocol for TMS, which involves seven steps: (1) research question, (2) literature review, (3) model space, (4) coverage, (5) computation, (6) descriptive analysis, and (7) multivariate analysis. The penultimate section (TBA) will provide a short replication exercise. The concluding section summarizes the approach and the findings along with some of the implications of TMS and plans for further development.

1 Research Question

Our protocol begins by identifying a research question centered on the impact of a causal factor (X) on an outcome (Y). Since the goal is to represent the scholarly literature on a topic, one must be attentive to how a question has been approached by previous studies. This is presumably often the most important consideration; but it is not the only consideration.

One must also consider how narrowly or broadly to frame the research question, as this determines the scope-conditions of the analysis. A narrowly defined research question (e.g., “Does democracy as measured by [Cheibub, Gandhi, and Vreeland \(2010\)](#)’s Democracy-Dictatorship index affect the first-difference of per capita log GDP as measured by [Feenstra, Inklaar, and Timmer \(2015\)](#)’s Penn World Tables?”)

will be easy to implement, while offering a modest payoff.

Of critical importance is the assumption that inputs and outputs of theoretical interest measure the same concepts. This is not always self-evident, and some conceptual choices may depend upon the theory that connects X to Y . Consider, once again, the relationship between democracy and growth. Some conceptualizations of democracy (say, particular deliberative notions of democracy) might not be expected to influence growth, and thus fall outside a plausible definition of democracy (for this particular research question).

When framing a question, one frequently encounters tradeoffs between narrow and broad conceptualizations. Small, tightly bounded research questions are more tractable, but also less helpful if the goal is to cumulate knowledge across many studies. Large, diffusely bounded research questions are harder to implement and more difficult to interpret, but – if successful – may be immensely valuable in cumulating results across a large field.

For this initial foray, we choose a middle path – large enough to encompass a sizeable literature but small enough to yield comparable results: *Does democracy cause growth?*

2 Literature

Having identified a research question, it is important to canvas the literature to see how that question has been addressed in the existing literature. The goal, more specifically, is to identify measures of X and Y , their functional forms, potential confounders, estimators, scope-conditions – any and all modeling choices that could help to delineate the TMS. Researchers should also take note of the frequency of these elements in the set of previous studies, distinguishing those that are widely embraced from those that are idiosyncratic.

For less studied topics, it may be necessary to examine studies adjacent to the question of theoretical interest in order to gain intuition about relevant modeling choices. Methodological studies may be particularly relevant. In our case, however, it seems reasonable to limit ourselves to studies explicitly focused on democracy and growth, which are numerous.

To canvas these studies, we draw on two recent comprehensive surveys of democracy and governance, one focused on the 2000-2020 period ([Gerring, Knutsen, and Berge, 2022](#)) and the other on the 2000-2019 period ([Colagrossi, Rossignoli, and Mag-](#)

[gioni, 2020](#)). Detailed appendices capture the methodological features of each study included in these compendiums, allowing us to gauge the frequency of each model choice.

3 Model Space

After reviewing the model choices revealed by a scholarly literature, the researcher(s) must assemble the elements into what we refer to as the "Total Model Space." To reiterate, TMS represents all those models (or designs) that could plausibly be utilized to estimate the impact of X on Y . This includes the many forking paths involved in research on a topic ([Gelman and Loken, 2014](#)).

The concept of *plausibility* is intentionally vague. We recognize that plausibility is a matter of degree and that researchers often disagree about which designs are plausible. This is especially true for the analysis of observational data, where there are usually many ways to estimate the impact of X on Y and many threats to inference to account for.

Our point of departure is to apply a rather liberal understanding – or a low bar – when delineating what the plausible specifications are. Our goal is not to start out by imposing our own judgments on these matters (and thus restricting the analysis from the outset), but rather to let the TMS reflect the judgments of (the various) experts who have spent their careers investigating a particular question. A strong clue to their intuitions lies in the published work they have (collectively) produced; hence, the central importance of a review of that literature in our protocol.

However, we cannot simply regurgitate the designs produced in published work without falling prey to potential problems of bias, as discussed at the outset. Rather than viewing models as settled destinations, we view them as clues to the paths that were, and were not, taken. In effect, we reverse-engineer the research process, treating published work as the residue of a set of decisions that are encoded in the eventual design. By examining all published work, we introduce even greater variety in these designs. And by taking the union of all design options across all published studies, we are able to identify a first-cut edition of the TMS. This entails culling the models of published work for all possible combinations of (a) measures of X , (b) measures of Y , (c) background covariates, (d) samples, and (e) estimators.

A second cut of the TMS aims to pare this model space in order to attain greater plausibility. After all, not every combination is plausible (even when we apply a

low bar), or even feasible. Specifically, one must be careful to exclude combinations of factors that are illogical or nonsensical, e.g., the inclusion of static background covariates in a unit fixed-effects analysis. One should also exclude specifications that are very difficult to defend and seldom employed, e.g., because a background factor is regarded as post-treatment. Occasionally, one must mandate the inclusion of a background variable in all models, e.g., if doing so is necessary in order for causal estimates to be meaningful. Careful consideration of the research question and causal relationships at play is required in this step, and all choices and restrictions should be explicitly documented for the sake of transparency.¹

If the first-cut TMS imposes excessive computational demands, further restrictions may be added in order to reduce the model space. For example, instead of treating background variables individually they may be randomly assigned into groups of three or four. By randomizing group assignments with each iteration one can still arrive at assessments of each variable’s impact on the estimation of bX , while vastly reducing computational demands.

Likewise, where samples vary in small ways across studies it makes no sense to treat each sample as a separate model choice. Instead, one must distinguish among sample *types* that have some resonance in the literature or theoretical rationale (e.g., OECD/non-OECD countries, 19th century/20th century). Each of these sample-types is treated as a separate modeling choice.

We reiterate that – in order to stay true to the literature – restrictions of the first-cut TMS should be invoked sparingly; they should rest upon priors well established by existing studies; and they should be noted explicitly. Anything defined out of TMS cannot be subjected to empirical tests and therefore rests on the author’s a priori assumptions. If they operation with one or more strong assumptions, the entire exercise rests on a weak reed. Insofar as the protocol is intended to aid knowledge cumulation, the model space must be defined broadly – this is why the TMS should include all *plausible* models, even if some are inevitably more plausible than others.

Following these general guidelines, we arrive at a TMS for democracy and growth, summarized in Table 1.

Time-frames are divided into two periods: historical (1800-2019) and contemporary (1960-2019). This distinction is important not only because it characterizes the field of study, but also because many of the background covariates commonly included

¹Researchers might also assess robustness by re-running the TMS with and without specific restrictions when it is unclear whether or not the omitted specifications qualify as plausible.

Table 1. TMS for Democracy and Growth

Modeling choice	Options
Sample construction	
Timeframe	2: Historical (1800-2019); Contemporary (1960-2019)
Country scope	3: All; Developing only; Only >1m population
Imputation	2: Non-imputed; Imputed using Amelia II with m=10
Dependent variable	
Source of DV	4: Fariss et al.; Maddison; PWT; WDI
Form of DV	2: Level Logged GDP PC; First-difference GDP PC
Forward of DV	4: No forward; 1 year; 5 years; 10 years
Independent variable	
Source of IndV	8: PolityV; V-Dem Polyarchy; V-Dem Liberal Dem.; Lexical; UDS; Polyarchy from observables; BMR; DD (expanded)
Form of IndV	2: Level; Stock with 5pct. depreciation rate
Model design	
Lags of Logged GDP PC	3: One-year lag; One- and two-year lag; One-, two-, and three- year lag
Time correction	3: None; yearly dummies; time trends
Country dummies	2: No; Yes
VCOV structure	4: IID; Hetero; Clustered; Twoway
Controls	
Controls	817: No controls; All 816 combinations of 3 of the following 18 controls: life expectancy, economic freedom, fertility, education, ethnolinguistic frac., logged gov. spending, logged investment, logged population, population growth, logged trade openness, latitude, inequality, logged inflation, international conflict, internal conflict, legal origin, logged natural resources, logged terms of trade
Restrictions	
• No models with logged GDP PC as dependent variable without time correction. • Only twoway SEs in models with time correction and country dummies. • No static controls (latitude, fractionalization) in models with country dummies.	
Note. See Table A1 for references and detailed indicators.	

in crossnational regressions are available only for the contemporary period.

Three country samples are recognized: (a) global, (b) developing world, and (c) non-micro states (at least one million inhabitants). This follows the literature, where distinctions are often drawn among these sub-samples.

In addition to these natural samples, we impute missing observations to construct imputed datasets, offering additional tests, as described in the next section.

Economic growth is measured in various ways in the literature. To represent this variety, we use four indicators: (a) a latent variable measure of GDP that draws information from the most widely used indicators (Fariss et al., 2022), (b) an updated version of the canonical historical dataset compiled some years ago by Angus Maddison (Bolt and van Zanden, 2020), (c) a measure compiled by the World Bank as part of the World Development Indicators (The World Bank, 2023), and (d) a measure drawn from the Penn World Tables (Feenstra, Inklaar, and Timmer, 2015). The two former measures have long time series, enabling us to use them both for historical and contemporary samples, whereas the two latter are only measured for recent decades.

Following the literature, growth is operationalized as a *level* variable, i.e., per capita GDP log (accompanied by a lagged dependent variable on the right side) or as a *first-difference* variable (percentage change in per capita GDP log). These outcomes are (a) measured in the same year, or forward lagged by (b) one year, (c) five years, or (d) ten years.²

Democracy is measured by several widely used indices along with a few additions that we regard as safeguards against measurement error.³ The resulting panel includes Polity2 from the Polity V dataset (Marshall and Gurr, 2020), the Electoral democracy index (or Polyarchy) from the V-Dem project (Coppedge et al., 2022), a version of the same index constructed from observables (Weitzel et al., 2024), the Liberal democracy index from V-Dem (Coppedge et al., 2022), the Lexical index of electoral democracy (Skaaning, Gerring, and Bartusevičius, 2015), the Unified Democracy Score (Pemstein, Meserve, and Melton, 2010), BMR (Boix, Miller, and Rosato, 2012), and Democracy-Dictatorship, with newly expanded coverage (Bjørnskov and Rode, 2019). Each index is also calculated as a stock variable (Gerring et al., 2005) with a five percent annual depreciation rate.

Per capita GDP (log) is introduced as a covariate on the right side with (a) one

²A forward lag for the outcome means that all independent variables are effectively backward-lagged.

³The one notable exclusion from our canvas of commonly used indices is Freedom House, whose coverage extends back only to the 1970s.

lag, (b) one and two lags, or (c) one, two, and three lags. Time-trend controls include (a) annual dummies, (b) a linear time trend, or (c) none. (The latter option is bypassed for models with a *level* measure of GDP, as this outcome is heavily trended and excluding time controls would open up for severe omitted variable bias and autocorrelation issues.) Models (a) may, or (b) may not, include country dummies. Standard errors are calculated as (a) IID, (b) heteroskedastic, (c) clustered by country, or (d) two-way (clustered by country and year).

Eighteen background covariates, which are assembled based on a review of the most common controls in the literature (drawing on information from [Gerring, Knutsen, and Berge, 2022](#)) and listed in Table 1, are tested in combinations of three, generating 816 permutations. By grouping the covariates, we reduce computational time while still obtaining information about which variables, or combinations of variables, impact the estimate of democracy on growth, as discussed. We also add one option with no covariates. Static covariates are excluded from models with country dummies.

The TMS that results from this set of permutations is extensive. Although not exhaustive, it presumably represents most of the commonly employed models in this wide-ranging literature as well as many other hitherto untested (but plausible) specifications.

4 Coverage

In a research setting with a modest number of units (namely countries), data coverage is critical. We propose several strategies to mitigate the problem of limited coverage.

First, if two or more background variables are highly correlated and conceptually similar, we retain the variable with greatest coverage. This enhances coverage – due to fewer missing observations from listwise deletion – while mitigating collinearity and computational demands, with minimal information loss. For example, the most common control variable across studies in our literature review relates to education and human capital (see Table A2). Some recent articles use complex indices such as the World Bank’s Human Capital Index ([The World Bank, 2023](#)) to operationalize the concept, but these measures are focused on recent years and cover a limited subset of countries. Therefore, we instead use average years of schooling for citizens 15 and over, from Clio Infra ([van Leeuwen and van Leeuwen-Li, 2015](#)), which is widely used in previous studies, provides historical coverage back to 1820, extensive cross-country

coverage, and is highly correlated with the Human Capital Index (Pearson’s $r=0.87$). Similar pragmatic considerations are made when operationalizing other controls.

Second, where possible, coverage for included variables is extended by choosing updated or improved sources, or by additional coding. In this fashion, we for example include a new dataset (Bjørnskov and Rode, 2019) that extends the coverage of the original Democracy-Dictatorship index (Cheibub, Gandhi, and Vreeland, 2010).

Even so, coverage is likely to be uneven across the chosen variables, and our collection of data is no exception. If uncorrected, this will truncate the dataset in potentially severe or idiosyncratic ways due to list-wise deletion, introducing bias and reducing statistical power.

Accordingly, the "natural" dataset is subjected to multiple imputation so as to generate ten versions of a complete dataset. Imputation extends to all country-years for which there is at least one measure of X (democracy) and one measure of Y (GDP) in the natural dataset. Following standard imputation procedures, estimates are averaged across these datasets and standard errors are adjusted (i.e., imputation-corrected) to reflect variability across the imputed observations. We use the Amelia II imputation program (Honaker, King, and Blackwell, 2011).

5 Computation

The introduction of myriad design features – inputs and outputs, functional forms, specifications, estimators, temporal units of analysis, and samples – and myriad combinations within and across features is considerably more ambitious than a standard set of robustness tests. As such, the TMS generates a non-trivial computational problem. In the case of democracy and growth, the TMS includes 211,175,424 separate models.

In order to improve the speed of the computations, we take advantage of R’s built-in parallelization capabilities (R Core Team, 2023), estimate the linear models using the state-of-the-art `fixest` R package (Bergé, 2018), and remotely deploy our computations using Amazon Web Services (AWS) EC2 cloud instances.⁴ On AWS, we spent approximately \$3.70 per one million models. Since all model runs are independent from each other, the process can be made arbitrarily fast by spreading the computations across cloud instances. (In our deployment, we used just one 196-threaded instance, *c6a.48xlarge*, and got to a speed of 31 minutes per one mil-

⁴<https://aws.amazon.com/ec2/>

lion models.) In future iterations we plan to provide comprehensive guidelines and open-source software to streamline the computation of other TMS applications.

Fortunately, we need not run every possible model in order to recover the information contained in that full set of models. Following prior work (Neumayer and Plümper, 2017; Rød, Knutsen, and Hegre, 2020; Steel, 2020, p. 673-5), we choose a random subset of analyses comprising two percent of the TMS for subsequent analyses, amounting to 4,613,422 models. The choice of two percent of models is informed by early tests and comparisons of how results change when we expand the percentage share of included models in the random sample. ⁵

6 Descriptive Analysis

As an overall measure of democracy’s impact on growth, we focus on t -values rather than coefficient estimates, as the latter are difficult to compare across the diversely constituted TMS, for example with some specifications using GDP per capita growth rate as outcome and others using (the very differently scaled) Ln GDP per capita. While standardized coefficients appear to assuage these problems, this expedient is problematic in the best of circumstances and especially so with a fixed-effect estimator, where sample variance is less relevant than through-time unit variance.

The first panel in Figure 1 reports the t -values for a random sub-sample comprising 2% of the TMS, as described above. It is tempting to regard the mean or median values (noted in blue and red, respectively) from this distribution as a “best guess” of democracy’s impact on growth. However, one must bear in mind that measures of central tendency are the product of an extremely inclusive collection of plausible models, each of which carries equal weight. Any change to the definition of plausibility is likely to impact the mean and median.

Nonetheless, the agnostic fashion by which model-choices are made in the TMS – essentially, a union of all possible model-choices registered in published research, with some exceptions in the form of our limited restrictions – offers ground for reflection. If the union of all plausible choices is starkly at variance with actual choices made in published work, one might wonder about the latter’s viability. In particular, one might suspect that the forks taken along the garden path are contrived to culminate at a preferred exit.

⁵When estimating our main specification in Table 2 with two separate one-percent samples of the TMS leads to coefficients with a correlation of 0.999.

With this in mind, it is fruitful to compare the results of the TMS to the t -values reported in published studies. To do so, we analyze results from two recent compilations (Gerring, Knutsen, and Berge, 2022; Colagrossi, Rossignoli, and Maggioni, 2020). This produces a comprehensive sample of work focused on democracy and growth, published between 2000 and 2019.

The second panel in Figure 1 includes results from all tests reported in the text or appendices of 81 studies, a total of 1,136 tests (collected by Colagrossi, Rossignoli, and Maggioni, 2020). The third panel focuses on the benchmark models from these studies, a total of 75 tests (collected by Gerring, Knutsen, and Berge, 2022).⁶

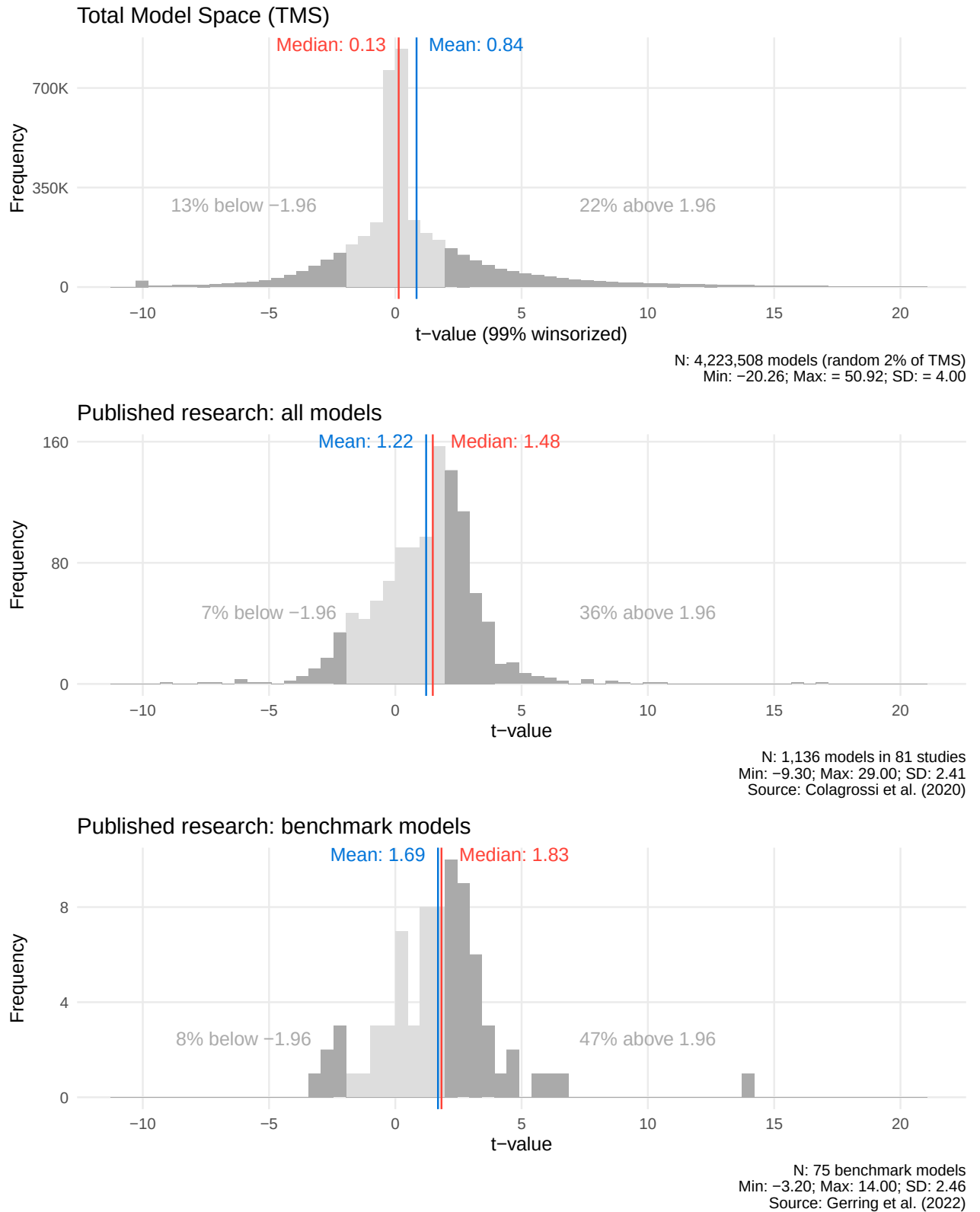
Across these three distributions, the TMS has the highest variance – a reflection of the agnostic approach we have taken to model choice. Accordingly, findings from published work fit within the bounds of the TMS.

The mean and median values from all three distributions are positive. However, they are close to zero for the TMS, higher for all tests reported in published studies, and higher still – indeed, close to conventional thresholds of statistical significance – for benchmark models from the published studies. Likewise, while only 22% of the models in the TMS lie above the 1.96 threshold, 36% and 47% (respectively) lie above the threshold in results culled from published research.

Published research thus suggests a considerably more optimistic picture of democracy’s impact on growth than the TMS. The contrast is especially stark if one focuses on benchmark models – which are, of course, authors’ preferred models and the results usually regarded as containing the central findings of published studies.

⁶In a future iteration, we expect to combine these two sources and code additional studies from 2020 onward.

Figure 1. Distributions of t-values



In addition to considering these central tendencies, one may compare the distribution of t-values around conventional cutoffs for statistical significance. The TMS histogram offers a smooth distribution with no apparent irregularities. All tests garnered from published studies reveal a mode lies just to the left of the 1.96 threshold and to the right of the more permissive 1.65 threshold. The benchmark models from these published studies reveal a mode lying just above the positive cutoff (1.96), and another mode just below the negative threshold (-1.96).

In comparing these distributions, it is difficult to avoid the conclusion that published work is subject to *p*-hacking or publication bias.

Finally, we should note that the left side of the histogram is more populated for the TMS than for published work. Negative t-values appear in 43.9% of the TMS, but in only 25.7% of all models and only 18.7% of benchmark models in published studies. This *could* reflect a pro-democracy bias among published studies. However, we reiterate that the TMS is a very permissive set of models, which may systematically differ from the published ones for various reasons. In the next section, we try to gain further insight into what types of models tend to produce particular types of results.

7 Multivariate Analysis

In addition to providing a benchmark by which to judge results from published studies, the TMS also offers insight into the data generating process. How do various modeling choices affect the estimand of interest (in this case, democracy’s impact on growth)?⁷

Outcomes of interest are the t-values from the TMS, displayed in the top panel of Figure 1. As previously, we draw on a random sample of 2%, constituting 4,613,422 democracy-growth models.

Two analyses of the data are summarized in Table 2. Both employ the same set of regressors, composed of dummies representing each modeling choice contained within the TMS (see Table 1). Despite the large number of right-side variables there is little threat of collinearity, as the sample is extremely large and the data evenly distributed across potential treatment conditions, minimizing correlations among predictors (see Figure A1). Analyses focus on point estimates, as traditional uncertainty measures make little sense with such a large number of observations (almost all relationships are statistically significant at conventional thresholds).

⁷A similar approach has been taken to understanding differences across text processing algorithms (Denny and Spirling, 2018).

Table 2. Results for random 2% of the TMS

	OLS Coef.	XGBoost Imp. (%)
GDP data source (ref: Fariss et al.)		16.46
WDI	1.30	9.36
Maddison	0.00	3.99
PWT	-0.88	3.11
Country fixed effects (ref: none)	0.32	13.36
Imputed sample (ref: no)	1.13	11.46
Standard errors (ref: iid)		10.62
Cluster	-0.79	8.60
Twoway	-0.88	1.75
Hetero	-0.01	0.27
Covariates (ref: life exp.)		9.98
Logged Terms Of Trade	-0.18	1.59
Education	-0.26	1.36
Inequality	-0.22	1.09
Economic Freedom	-0.39	1.05
Fertility	-0.32	0.93
Logged Inflation	-0.18	0.76
Logged Natural Resources	0.17	0.60
Legal Origin	0.12	0.56
Logged Population	0.10	0.34
Logged Trade Openness	0.02	0.34
Population Growth	0.09	0.31
Logged Gov. Spending	0.05	0.30
Logged Investment	0.12	0.29
Latitude	0.19	0.16
Ethnolinguistic Frac.	0.07	0.11
Internal Conflict	0.15	0.10
International Conflict	0.17	0.08
Growth: first-dif (ref: level)	-2.60	9.27
Countries (ref: all)		6.53
Developing	-0.54	4.88
Pop Over 1 Mil.	-0.39	1.65
Democracy index source (ref: Polity V)		6.40
V-Dem Liberal Dem	0.29	1.17
Lexical	0.08	0.93
UDS	0.13	0.92
Polyarchy (From Observables)	0.24	0.90
Dd (Expanded)	0.04	0.89
BMR	-0.05	0.79
Polyarchy (V-Dem)	0.14	0.79
Democracy stock (ref: level)	-0.47	5.32
Lag structure (ref: none)		4.27
10 Years	0.29	2.70
5 Years	0.15	1.14
1 Year	-0.16	0.43
Time-frame: Contemporary (ref: Historical)	-0.39	3.01
Trend correction (ref: none)		2.71
Annual Dummies	-0.78	2.18
Linear Trend	-0.20	0.53
Lags of GDP (ref: 1-year lag)		0.61
1-, 2-, And 3- Year Lag	-0.04	0.34
1- And 2- Year Lag	0.03	0.27
Out-of-sample RMSE	3.60	0.60
Adjusted R2	0.18	
Num. obs.	4,223,508	4,223,508

Model 1 is an ordinary least squares (OLS) analysis. Because all right-side variables are composed of dummies, coefficients are directly comparable – though one must bear in mind that estimated effects are judged against a reference category (listed in parentheses).

The sign of the coefficient indicates the impact of a factor across all models in the TMS. If positive (negative), the design choice enhances (diminishes) the estimated impact of democracy on growth.

The model fit statistic shows that less than twenty percent of the variability in results is accountable to model choices in a simple additive model. This, in turn, suggests that there may be important interactions among model choices.

Model 2 shows the importance of each variable in an Extreme Gradient Boosting (XGBoost) regression (Chen and Guestrin, 2016). Similar to a random forest, XGBoost operates with a tree structure, allowing it to test variable combinations. Previous research in political science has demonstrated that tree-based methods can detect complex non-linearities and interactions in applied settings (Montgomery and Olivella, 2018).

Following standard practice, we use cross-validation to choose the optimal depth of the trees. In the table, we follow Chen et al. (2023), reporting variable importance as the share of total fit gain that the variable adds across all splits. By design, percentages add up to 100%, making interpretation straightforward.

XGBoost variable importance values from Model 2 are (weakly) rank-correlated with the absolute values of coefficients from OLS ($\rho = 0.72$). In order to assess model performance, we examine a holdout set constructed from a different random sample of the full TMS, with 211,175 models (0.1% of the TMS). In this holdout set, which offers an opportunity for out-of-sample validation, OLS obtains a root mean squared error (RMSE) of 3.60, while XGBoost is at 0.60—a substantial improvement in model fit. This corroborates our intuition that the tree-based method is capturing something beyond what is possible using simple additive models.

A supplementary model, displayed in Appendix ??, presents results from a LASSO regression (Tibshirani, 1996).⁸ By construction, the correlation between coefficients in the OLS and Lasso models is nearly perfect, at 0.99. Naturally, model-fit is improved. However, since little additional information is provided by LASSO relative to OLS

⁸We select a suitable shrinkage hyperparameter via cross-validation, using the implementation from Friedman, Tibshirani, and Hastie (2010). Most coefficients shrink towards zero and three design features are dropped entirely: one source of democracy scores, one background covariate, and the heteroscedasticity-robust correction for standard errors.

and XGBoost, these results are relegated to an appendix.

7.1 What matters

In Table 2, variables are analyzed separately and also aggregated into categories so as to offer a sense of how much of the variability each category explains in the XGBoost model. More important categories are listed first. Overall, it appears that six design choices have an especially strong influence on crossnational tests of democracy and growth.

The most important feature is the outcome measure and its data source. Regression tests reveal a much more optimistic picture of democracy’s effect on economic performance when the measure of growth is drawn from the WDI dataset. Other data sources – including Fariss, Maddison, and PWT – offer a more pessimistic assessment. This does not appear to be a product of how GDP is operationalized, e.g., local currency or US dollars, current exchange rates or purchasing power parity, natural or inflation-adjusted values.

Next in importance is the choice of whether or not to include country dummies. The aggregate effect of this design feature is modestly positive, signaled by an OLS coefficient of 0.32, reflecting that, on average, adding country dummies enhances democracy’s estimated effect on growth. However, there appears to be a good deal of variability across models when it comes to exactly how the inclusion of country-fixed effects affects results, signalled by high variable importance in the XGBoost model. We conclude that the inclusion of country dummies has a substantial impact on the relationship of democracy to growth but that the direction of the impact depends upon other modeling choices.

The third most important feature concerns the imputation of missing data. This feature accounts for 11.5% of the total model gain across all folds. Imputed samples generally show a stronger (positive) relationship between democracy and growth. Model 1 indicates that the average t -value for the democracy coefficient is one point higher in imputed samples than in non-imputed samples.

There are two complementary explanations that help account for this substantial effect of using imputed samples. First, the larger sample afforded by imputation reduces standard errors, all things equal, which “mechanically” enhances (absolute) t -values. Second, there are likely substantive differences across natural and imputed samples. Consider that missingness is likely to affect small, poorly governed countries. If many of these are autocracies with weak economic performance it is easy to see how

the inclusion of imputed data would lead to a more optimistic picture of democracy’s impact on growth (Knutsen, 2021).

A fourth design feature of importance is how standard errors are calculated. The baseline reference point is the classical assumption of variables that are independent and identically distributed (iid). When corrections are introduced to account for probable violations of this assumption, the relationship between democracy and growth attenuates. This is what one would expect mechanically, as corrections typically reduce SEs and larger SEs decrease t-values. However, it is informative to know which options have the greatest impact on the substantive results. This turns out to be the decision about whether or not to cluster standard errors by country. Accounting for serial autocorrelation in this manner has a strongly negative impact on t-values pertaining to the democracy-growth relationship.

The selection of control variables has a relatively modest influence on estimates of democracy’s impact on growth. Collectively, the inclusion or exclusion of seventeen covariates account for only ten percent of the variability in the XGBoost model. Considered individually, some of these variables enhance the estimated relationship between democracy and growth while other diminish it. But none have an especially strong impact.⁹

To our surprise – especially given the lack of explicit attention placed on this choice in previous studies – the functional form of the outcome has a substantial impact on democracy’s estimated relationship to growth. This decision alone accounts for over nine percent of the variability in the XGBoost model. Moreover, its impact is almost entirely in one direction. When economic performance is measured by the one-period change in per capita GDP (log), approximating the annual growth rate in GDP per capita, the average *t*-value associated with democracy’s impact on growth diminishes by 2.6 points relative to the baseline. The baseline is the alternate functional form, using a “level” measure of per capita GDP (log). (All specifications include per capital GDP (log) as a covariate on the right side of the model.)

Proponents of level specifications may point out that the higher average t-values could partly reflect lower average standard errors stemming from the much lower variance (and related measurement error) for GDP per capita. Proponents of first-difference specifications may point out that level measures enhance the possibility that autocorrelation will artificially inflate t-values (though it should be noted that we

⁹Please note that many of these covariates might be considered post-treatment, and controlling for them might therefore bias the estimated (total) effect of democracy. Note also that static variables such as Latitude are included only in models without country fixed-effects.

omit level specifications without any time trend controls). In any case, this difference in model choice, which is not widely noted in the literature, clearly deserves more attention. It may help explain why economists, who are more likely to employ level measures of growth, also tend to be more optimistic than political scientists about the relationship between democracy and growth (e.g., [Acemoglu et al. \(2019\)](#)).

Other model choices have less impact on the estimated relationship between democracy and growth across the TMS, whether understood in an additive model (OLS) or a branching model (XTBoost). This includes the country sample, where we compared a global (unrestricted) sample to samples focused on the developing world or sizeable countries (excluding countries with populations of less than one million). It includes the sample time-frame, where we compared the contemporary era (1960-2020) to the entire modern era (1800-2020). It includes the chosen measure of democracy, where we compared eight widely used indices. It includes the treatment of democracy as a "level" or "stock" measure (incorporating a country's regime history). It includes lag structure, right-side lags of per capita GDP, and trend corrections such as annual dummies and a linear time trend.

Together, these analyses are immensely informative insofar as they summarize at a glance the practical impact of all the (plausible) design choices that inform the scholarly literature on this topic. Specifically, they allow us to distinguish high-impact design choices from low-impact design choices, which one may regard (individually) as relatively trivial.

8 Replication

In this section, we plan to replicate a study. Because of its prominence we will probably focus on [Acemoglu et al. \(2019\)](#). We will begin by adding their chosen measure of democracy to our TMS and seeing if we can approximate their benchmark model. (Ultimately, we hope to do the same for other benchmark models culled from the published literature.)

With this as the point of departure, we will introduce changes to the benchmark in the usual fashion of robustness tests – changing one feature at a time while keeping other features constant. This generates a series of t-values that can be displayed in a histogram and compared to the histogram from the TMS in Figure 1.

Our expectation is that robustness tests reveal a restricted, and maybe highly skewed picture of reality insofar as the point of departure – the benchmark model

– limits the potential model space. If the benchmark model is optimistic, small departures from the benchmark are also likely to be optimistic.

Additionally, we anticipate that the marginal effect of a particular change to a specific benchmark model differs from the mean effect of the same model change across the much wider set of models contained in the TMS. In other words, the effects of making certain modeling choices may be heterogeneous and contingent on a host of other modeling choices. If so, this illustrates another limitation of the usual approach to robustness.

Of course, it is still valuable to know which departures have the greatest impact on the benchmark findings. And the TMS offers a handy way of introducing these and thus provides a machinery for robustness testing that can be easily deployed (so long as the necessary features of the benchmark are contained in the TMS or can be easily added). We recommend histograms of t-values (or coefficients) that indicate the benchmark result along with all the other results generated by the robustness tests as a handy presentational device.

9 Discussion

In this study, we have outlined an approach to knowledge cumulation with observational data, which we dub total model space (TMS). Applying this approach to the time-honored question of democracy’s impact on growth, we outlined a protocol with seven steps: (1) specifying the research question, (2) reviewing the literature, (3) defining the TMS, (4) expanding coverage, (5) computing models within the TMS, and finally, analyzing the impact of various modeling choices descriptively (6) and in a multivariate analysis (7). In this final section, we discuss the payoffs, implications, and future course of this novel approach in three parts: *credibility*, *diagnostics*, and *further uses*.

9.1 Credibility

Our point of departure is the assumption that analyzing literatures is more credible than analyzing individual studies. Even so, one cannot accept published results, or even some aggregate of all of these results, at face value, as they are subject to potential biases arising from the research- and publication process. One must also examine the set of plausible designs on a given subject from which published designs are drawn.

This set of plausible designs composes the total model space. We have shown how the TMS can be recovered by unraveling the set of choices embedded in published benchmark models. One can then compare the set of results produced by all *plausible* models with those recorded in *published* models. As it happens, the former are, on average, considerably less optimistic than the latter about the contribution of democracy to growth.

We do not have grounds to assert with confidence that the mean (or median) value across the t-values generated from the TMS is a better guess of the true relationship between X and Y than the mean (or median) from a distribution of t-values generated by published work. But the disjuncture does give one pause. If plausible model choices were substantively neutral with respect to X 's impact on Y , one would expect the two distributions to exhibit similar central tendencies. They do not. Moreover, the peculiar shape of the distribution of t-values for published work on democracy and growth is consistent with oft-discussed patterns of p-hacking (by researchers, or similar biases among reviewers or editors downstream in the publication process). Differences in the distributions between TMS results and published results are also consistent with what one might call normative hacking – in this case, the presumed preference for democracy to have a positive impact on growth – although we must note that we do not have any direct evidence that such a bias has contributed substantially to the difference. There might be other explanations for this difference.

In future research, we plan to delve into the concept of *plausibility*. Instead of treating all models in the TMS as equally plausible, it may be possible to provide informed guesses about which models are more/less plausible. One approach relies on the literature, counting up the share of times a particular model choice is included among published studies, as we have for background covariates in Table A2. Yet, this approach is susceptible to different biases, since it reflects choices made in benchmark models (with the potential issues mentioned directly above related to p-hacking and normative hacking).

A better approach separates methodological judgments from substantive findings. One way to do this is to pose question about research design to experts who are familiar with the substantive research question along with methodologists who are familiar with the methodology. By-passing model choices that have minimal impact on the question (as revealed by our analysis of the TMS), such a survey can home in efficiently on a small handful of choices that are most consequential.

Whether inferred from published studies or from expert surveys, these method-

ological choices can then be adopted as priors in a Bayesian-style analysis, providing a methodologically informed adjustment to the TMS. This, in turn, might have credibility as a "best-guess" about democracy's impact on growth – and, for that matter, about any other question for which there is an established literature and a recognizable set of experts.

9.2 Diagnostics

In addition to an overall distribution, we have offered a variety of diagnostic tools to understand variability across t-values produced by the TMS. Having coded each design feature in a series of dummy variables, these are employed on the right side of regression models. OLS, Lasso, and XGBoost estimators tell a similar story, with a number of unexpected twists.

Some unexpected features that have received little attention in the literature – such as the choice between level and first-difference functional forms or between natural and imputed datasets – turn out to be quite important. Other features such as background covariates, the usual suspects when it comes to arguments about model choice, turn out to be less consequential. At least in one case – the "stock" measure of democracy – the anticipated effect of a design feature is the opposite of what authors have previously found (Gerring et al., 2005).

However, we need to know more about the gigantic black box generated by the 211,175,424 models in our full TMS. In future work, we plan to examine heterogeneity across this vast model space so as to better understand why some design choices register stronger effects than others, and what combinations of background features contribute to these effects. We expect that there is considerable heterogeneity hidden behind the millions of models within the TMS.

9.3 Further uses

Although the approach outlined in this paper may seem a good deal more complicated than the standard empirical social science paper, it is intended to establish a template that is easily adapted for use in other contexts. (To facilitate this, a complete dataset and library in R will be posted on Dataverse.)

Many efficiencies may be anticipated vis-a-vis the usual approach to crossnational research. Consider that authors spend an extraordinary amount of time compiling datasets and reviewing the literature on a subject. With a TMS dataset at hand, all that is needed is the addition of one or several new features of special interest,

e.g., a new measure of the input or output, a new background covariate, a new scope-condition, or a new estimator. This can be integrated with little effort into the existing TMS dataset. With small modifications to the code, a complete set of tests can be run, generating a new and further expanded TMS, which in turn paves the way for future research.

Efficiencies are presumably greatest when scholars are working on the same research question. Yet, because there is so much overlap in crossnational research – empirically, theoretically, and methodologically – one can anticipate efficiencies even when X or Y are different. For example, scholars interested in human development, conflict, or some other aspect of governance may use our dataset and accompanying code, crafted for democracy and growth, as a starting point. This lowers entry costs for those wishing to contribute with a new idea (or several) to an established line of research.

The envisioned protocol also provides a consistent and structured framework for bringing replication into the research process. Every new study must incorporate all previous analyses, or offer a compelling reason for their exclusion. Since it is fairly easy to do so – the code is already written and easy to adapt – no additional burden is placed on the author.

Equally important, with our approach it is difficult for authors to fish for desired findings by ignoring results that do not support their favored hypothesis. The TMS protocol requires a comprehensive cull of variables, functional forms, estimators, and scope-conditions.

Although it may not be immediately apparent, our protocol should be fully compatible with research of an inductive or deductive nature. Scholars with weak priors can probe multiple hunches without having to make a commitment. Scholars with strong priors can implement that theoretical expectation. If it involves the collection of new data, not heretofore available, that theoretical expectation might be pre-registered along with a detailed pre-analysis plan. In any case, the results of the favored hypothesis and research design may be compared and contrasted with other hypotheses and designs with little additional effort.

Eventually, we hope to create an online interface allowing end-users to choose design preferences from a menu of options – including all those discussed in this paper. The resulting t-value could be produced, showing its placement within the histogram of the entire TMS. In this fashion, the implications of model-choices could be tested easily and virtually instantaneously. In addition to providing a useful tool

for teaching, we hope that this might facilitate methodological reflection in a more considered and focused fashion.

References

- Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A Robinson. 2019. “Democracy does cause growth”. *Journal of political economy* 127 (1):47–100.
- Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. “Fractionalization”. *Journal of Economic Growth* 8 (2):155–194.
- Barbieri, Katherine, Omar M.G. Keshk, and Brian M. Pollins. 2009, October. “Trading data”. *Conflict Management and Peace Science* 26 (5):471–491.
- Barbieri, Katherine and Omar M. G. Keshk. 2016. “Correlates of war project trade data set codebook, version 4.0.”.
- Bergé, Laurent. 2018. “Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm”. *CREA Discussion Papers* (13).
- Bjørnskov, Christian and Martin Rode. 2019, February. “Regime types and regime change: A new dataset on democracy, coups, and political institutions”. *The Review of International Organizations* 15 (2):531–551.
- Boix, Carles, Michael Miller, and Sebastian Rosato. 2012, November. “A complete data set of political regimes, 1800–2007”. *Comparative Political Studies* 46 (12):1523–1554.
- Bolt, Jutta and Jan Luiten van Zanden. 2020. “Maddison style estimates of the evolution of the world economy. a new 2020 update.”.
- Brecke, Peter. 2015a. “Armed conflicts (internal) [dataset]”.
- Brecke, Peter. 2015b. “Armed conflicts (international) [dataset]”.
- Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. “Methods matter: P-hacking and publication bias in causal analysis in economics”. *American Economic Review* 110 (11):3634–60.
- Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland. 2010, August. “Democracy and dictatorship revisited”. *Public Choice* 143 (1-2):67–101.
- Chen, Tianqi and Carlos Guestrin. 2016. “Xgboost: A scalable tree boosting system”. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.
- Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, and Jiaming Yuan. 2023. *xgboost: Extreme Gradient Boosting*. R package version 1.7.5.1.

- Colagrossi, Marco, Domenico Rossignoli, and Mario A. Maggioni. 2020. “Does democracy cause growth? a meta-analysis (of 2000 regressions)”. *European Journal of Political Economy* 61 : 101824.
- Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2022. “V-dem codebook v12”.
- Davies, Shawn, Therése Pettersson, and Magnus Öberg. 2023. “Organized violence 1989–2022, and the return of conflict between states”. *Journal of Peace Research*.
- de Marchi, Scott, John Gerring, and Kyosuke Kikuta. 2023. “Deep inequalities”. Unpublished manuscript.
- de Zwart, Pim. 2015. “Inflation [dataset]”.
- Denny, Matthew J. and Arthur Spirling. 2018. “Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it”. *Political Analysis* 26 (2): 168–189.
- Durlauf, S. N., P. Johnson, and Jonathan Temple. 2005. “Growth econometrics”. In P. Aghion and S. N. Durlauf (Eds.), *Handbook of Economic Growth, Vol 1A*, pp. 555–677. Amsterdam: North-Holland.
- Fariss, Christopher J., Therese Anders, Jonathan N. Markowitz, and Miriam Barnum. 2022, February. “New estimates of over 500 years of historic GDP and population data”. *Journal of Conflict Resolution* 66 (3): 553–591.
- Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015, October. “The next generation of the penn world table”. *American Economic Review* 105 (10): 3150–3182. [version 10.01].
- Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. “Regularization paths for generalized linear models via coordinate descent”. *Journal of Statistical Software* 33 (1): 1–22.
- Gapminder. 2023. “Gapminder data”.
- Gelman, Andrew and Eric Loken. 2014. “The statistical crisis in science”. *American Scientist* 102 (6): 460–465.
- Gerring, John, Philip Bond, William T Barndt, and Carola Moreno. 2005. “Democracy and economic growth: A historical perspective”. *World Politics* 57 : 323–364.

- Gerring, John, Carl Henrik Knutsen, and Jonas Berge. 2022. “Does democracy matter?”. *Annual Review of Political Science* 25 : 357–375.
- Haber, Stephen and Victor Menaldo. 2011, January. “Do natural resources fuel authoritarianism? a reappraisal of the resource curse”. *American Political Science Review* 105 (1): 1–26.
- Harbom, Lotta, Erik Melander, and Peter Wallensteen. 2008, September. “Dyadic dimensions of armed conflict, 1946—2007”. *Journal of Peace Research* 45 (5): 697–710.
- Honaker, James, Gary King, and Matthew Blackwell. 2011. “Amelia II: A program for missing data”. *Journal of Statistical Software* 45 (7): 1–47.
- Ioannidis, John PA. 2005. “Why most published research findings are false”. *PLoS Medicine* 2 (8): e124.
- Kittel, Bernhard. 2006. “A crazy methodology? on the limits of macro-quantitative social science research”. *International sociology* 21 (5): 647–677.
- Knutsen, Carl Henrik. 2021. “A business case for democracy: regime type, growth, and growth volatility”. *Democratization* 28 (8): 1505–1524.
- La Porta, Rafael, Florencio Lopez-de Silanes, and Andrei Shleifer. 2008, May. “The economic consequences of legal origins”. *Journal of Economic Literature* 46 (2): 285–332.
- Magee, Christopher S. P. and John A. Does. 2015, June. “Reconsidering regime type and growth: Lies, dictatorships, and statistics”. *International Studies Quarterly* 59 (2): 223–237.
- Marquez, Xavier. 2016, March. “A Quick Method for Extending the Unified Democracy Scores”.
- Marshall, Monty G. and Ted Robert Gurr. 2020. “Polity5 Project: Political Regime Characteristics and Transitions, 1800-2018”.
- Moatsos, Michalis, Jan Luiten van Zanden, Joerg Baten, Peter Foldvari, and Bas van Leeuwen. 2015. “Income inequality [dataset]”.
- Montgomery, Jacob M. and Brendan Nyhan. 2010. “Bayesian model averaging: Theoretical developments and practical applications”. *Political Analysis* 18 (2): 245–270.
- Montgomery, Jacob M and Santiago Olivella. 2018. “Tree-based models for political science data”. *American Journal of Political Science* 62 (3): 729–744.
- Muñoz, John and Cristobal Young. 2018. “We ran 9 billion regressions: Eliminating false positives through computational model robustness”. *Sociological Methodology* 48 (1): 1–33.

- Neumayer, Eric and Thomas Plümper. 2017. *Robustness tests for quantitative research*. Cambridge University Press.
- Pemstein, Daniel, Stephen A. Meserve, and James Melton. 2010. “Democratic compromise: A latent variable analysis of ten measures of regime type”. *Political Analysis* 18 (4): 426–449.
- Pettersson, Therese. 2023. “Ucdp dyadic dataset codebook v23.1”.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rodrik, Dani. 2012. “Why we learn nothing from regressing economic growth on policies”. *Seoul Journal of Economics* 25 (2): 137–151.
- Rød, Espen Geelmuyden, Carl Henrik Knutsen, and Håvard Hegre. 2020. “The determinants of democracy: A sensitivity analysis”. *Public Choice* 185 (1): 87–111.
- Seawright, Jason. 2007. “Democracy and growth: A case study in failed causal inference”. In G. Munck (Ed.), *Regimes and democracy in Latin America: theories and methods*, pp. 179–98. Oxford: Oxford University Press.
- Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020, July. “Specification curve analysis”. *Nature Human Behaviour* 4 (11): 1208–1214.
- Skaaning, Svend-Erik, John Gerring, and Henrikas Bartusevičius. 2015. “A lexical index of electoral democracy”. *Comparative Political Studies* 48 (12): 1491–1525.
- Solt, Frederick. 2020. “Measuring income inequality across countries and over time: The standardized world income inequality database”. *Social Science Quarterly* 101 (3): 1183–1199.
- Steege, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. “Increasing transparency through a multiverse analysis”. *Perspectives on Psychological Science* 11 (5): 702–712.
- Steel, Mark FJ. 2020. “Model averaging and its use in economics”. *Journal of Economic Literature* 58 (3): 644–719.
- The World Bank. 2023. “World development indicators [dataset]”.
- Tibshirani, Robert. 1996. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–288.
- van Leeuwen, Bas and Jieli van Leeuwen-Li. 2015. “Average years of education [dataset]”.

- Weitzel, Daniel, John Gerring, Dan Pemstein, and Svend-Erik Skaaning. 2024. “Measuring electoral democracy with observables”. *American Journal of Political Science*.
- Young, Cristobal and Katherine Holsteen. 2017. “Model uncertainty and robustness: A computational framework for multimodel analysis”. *Sociological Methods & Research* 46 (1): 3–40.

A Appendix: Variables and Data Sources

All models that we estimate regress a measure of economic growth (Y) on a measure of democracy (X). Other variables are used as controls. Table [A1](#) describes all variables and their sources.

Appendix Table A1. Variables and sources

Variable	Indicator	Coverage	Source
Economic Development (and Growth)	Gross Domestic Product (per capita)	1800-2019	Fariss et al. (2022), latent version
	Gross Domestic Product (per capita)	1800-2018	Maddison (Bolt and van Zanden, 2020)
	Gross Domestic Product, PPP (per capita) (output-side)	1960-2019	PWT 10.01 (Feenstra, Inklaar, and Timmer, 2015)
	Gross Domestic Product (per capita)	1960-2019	WDI (The World Bank, 2023)
Democracy	Electoral democracy index	1800-2019	V-Dem (Coppedge et al., 2022)
	Liberal democracy index	1800-2019	V-Dem (Coppedge et al., 2022)
	Polity V, revised combined score	1800-2018	Marshall and Gurr (2020) drawn from V-Dem (Coppedge et al., 2022)
	Lexical Index of Electoral Democracy	1800-2019	Skaaning, Gerring, and Bartusevičius (2015) drawn from V-Dem (Coppedge et al., 2022)
	Unified Democracy Score (UDS)	1800-2019	Marquez (2016)'s update of Pemstein, Meserve, and Melton (2010)
	Electoral democracy (predicted from observables)	1800-2019	Weitzel et al. (2024)
	BMR index (dichotomous)	1800-2015	Boix, Miller, and Rosato (2012) drawn from V-Dem (Coppedge et al., 2022)
	Democracy-Dictatorship (dichotomous)	1960-2019	Bjørnskov and Rode (2019)'s update of Cheibub, Gandhi, and Vreeland (2010)
Education	Average years of schooling, 15 and over	1820-2019	Clio Infra (van Leeuwen and van Leeuwen-Li, 2015) drawn from V-Dem (Coppedge et al., 2022)
Investment	Share of gross capital formation at current PPPs	1960-2019	PWT (Feenstra, Inklaar, and Timmer, 2015)
Trade	(Exports + Imports) / GDP PPP	1870-2014	COW (Barbieri and Keshk, 2016; Barbieri, Keshk, and Pollins, 2009) for exports and imports; Fariss et al. (2022) for GDP
Government spending	Government consumption as share of GDP at current PPP	1960-2019	PWT (Feenstra, Inklaar, and Timmer, 2015)

Population (and Growth)	Population	1800-2019	Fariss et al. (2022)
Fractionalization	Ethnic + Language Fractionalization (PCA 1st component)	-	Alesina et al. (2003)
Life expectancy	Life Expectancy at Birth	1800-2019	Gapminder (2023) drawn from V-Dem (Coppedge et al., 2022)
Fertility	Babies per woman (total fertility rate)	1800-2019	Gapminder (2023)
Tropical climate	Absolute latitude (capital)	1800-2019 (static)	de Marchi, Gerring, and Kikuta (2023)
Economic freedom	Property rights - Do citizens enjoy the right to private property?	1800-2019	V-Dem (Coppedge et al., 2022)
Inflation	Inflation, annual percentage change	1800-2010	Clio Infra (de Zwart, 2015)
	Inflation, consumer prices (%)	1960-2019	The World Bank (2023)
Inequality	Gini index of inequality, equivalized, household disposable	1960-2019	Solt (2020)
	Gross household income Gini	1820-2000	Clio Infra (Moatsos et al., 2015)
International conflict	Armed conflict, international	1800-2000	Clio Infra (Brecke, 2015b)
	Armed conflict (extrasystemic, interstate)	1960-2019	UCDP (Pettersson, 2023 ; Harbom, Melander, and Wallensteen, 2008 ; Davies, Pettersson, and Öberg, 2023)
Internal conflict	Armed conflict, internal	1800-2000	Clio Infra Brecke (2015a)
	Armed conflict (intrastate, internationalized intrastate)	1960-2019	UCDP (Pettersson, 2023 ; Harbom, Melander, and Wallensteen, 2008 ; Davies, Pettersson, and Öberg, 2023)
Resource wealth	Petroleum, coal, natural gas, and metals production per capita	1900-2006	Haber and Menaldo (2011)
Terms of trade	Total natural resources rents (% of GDP)	1970-2019	The World Bank (2023)
	Net barter terms of trade index	1980-2019	The World Bank (2023)
Legal Origin: five traditions (dummies)	Legal Origin	1800-2019 (static)	La Porta, Lopez-de Silanes, and Shleifer (2008)

In order to select these covariates, we examined 80 articles which test for the relationship between democracy and economic growth, as identified by [Gerring, Knutsen, and Berge \(2022\)](#). Table A2 shows the 22 control variables that appear in at least three of these articles. We do not include data for currency exchange or regime durability.

Appendix Table A2. Appearances of controls in the literature

Control variable	Appearances (out of 80)
Education and Human Capital	34
Investment	22
Trade	15
Government Spending	14
Population Growth	12
Ethnic Fractionalization	9
Life Expectancy	9
Population	7
Fertility	6
Tropical Climate	6
Economic Freedom	6
Inflation	5
Currency Exchange	5
Property Rights	4
Inequality	4
International Conflict	4
Resource Wealth	4
Terms of Trade	4
Institutional Quality	3
Political Instability	3
Regime Durability	3
Legal Origin	3

Appendix Table A3. Descriptive statistics for variables in models, before imputation

	Non-missing	N countries	N years	Year range	Mean	SD	Min	Max
Growth: Logged GDP PC (Fariss et al.)	94.2%	222	220	1800-2019	8.21	1.24	5.67	14.25
Growth: Logged GDP PC (Maddison)	80.6%	168	219	1800-2018	15.09	1.12	12.59	18.87
Growth: Logged GDP PC (PWT)	33.7%	174	60	1960-2019	8.79	1.20	5.37	12.62
Growth: Logged GDP PC (WDI)	36.7%	197	63	1960-2022	7.64	1.74	2.55	12.36
Growth: first difference of GDP PC (Fariss et al.)	93.4%	222	219	1801-2019	1.29	3.35	-51.46	67.00
Growth: first difference of GDP PC (Maddison)	80.0%	168	218	1801-2018	1.26	6.06	-72.95	173.89
Growth: first difference of GDP PC (PWT)	33.1%	174	59	1961-2019	2.58	9.96	-82.41	340.61
Growth: first difference of GDP PC (WDI)	36.0%	197	62	1961-2022	6.50	16.47	-99.78	398.70
Democracy: V-Dem Electoral democracy	80.1%	192	222	1800-2021	0.30	0.27	0.01	0.93
Democracy: V-Dem Liberal democracy	77.9%	191	222	1800-2021	0.24	0.24	0.01	0.90
Democracy: Lexical scores	69.1%	190	220	1800-2019	2.77	2.30	0.00	6.00
Democracy: PolityV	62.8%	179	219	1800-2018	-0.48	7.08	-88.00	10.00
Democracy: Electoral democracy from observables	92.2%	214	220	1800-2019	0.28	0.25	0.02	0.92
Democracy: Unified Democracy Scores (UDS)	96.1%	226	223	1800-2022	0.20	0.93	-1.76	3.13
Democracy: BMR (dichotomous)	66.2%	186	223	1800-2022	0.33	0.47	0.00	1.00
Democracy: DD (dichotomous)	49.5%	195	73	1950-2022	0.43	0.50	0.00	1.00
Education: Average years of schooling, 15 and over	51.4%	137	202	1820-2021	4.80	3.50	0.03	13.61
Investment: Logged Share of gross capital formation at current PPPs	33.7%	174	60	1960-2019	3.42	0.34	0.00	5.79
Trade: Logged (Exports + Imports) / GDP PPP	45.7%	177	145	1870-2014	1.95	1.20	0.00	7.17
Gov. spending: Logged gov. consumption as share of GDP at current PPP	33.7%	174	60	1960-2019	2.90	0.50	0.42	5.36
Logged population	94.2%	222	220	1800-2019	15.16	1.96	7.73	21.12
Population growth	93.4%	222	219	1801-2019	1.31	1.62	-25.44	59.87
Ethnic + Language Fractionalization (PCA 1st component)	94.1%	194	223	1800-2022	-0.04	0.33	-0.58	0.73
Life Expectancy at Birth	78.5%	173	222	1800-2021	51.32	16.90	1.50	85.10
Fertility: Babies per woman	88.5%	188	220	1800-2019	4.86	1.86	0.83	8.87
Absolute latitude (capital)	82.2%	202	221	1800-2020	0.31	0.19	0.00	0.72
Property rights: Do citizens enjoy the right to private property?	81.8%	192	222	1800-2021	0.50	0.29	0.00	0.97
Logged inflation	50.4%	189	223	1800-2022	5.35	0.25	0.00	26.08
Inequality: Gini	43.4%	188	202	1820-2021	40.79	8.65	17.90	67.30
International conflict	81.6%	185	220	1800-2019	0.10	0.31	0.00	1.00
Internal conflict	81.6%	185	220	1800-2019	0.11	0.31	0.00	1.00
Nat res: Logged total natural resources rents (% of GDP)	32.0%	197	52	1970-2021	1.38	1.18	0.00	4.50
Nat res: Logged Petroleum, coal, natural gas, and metals production pc	39.0%	163	107	1900-2006	3.27	2.56	0.00	11.30
Terms of trade: Logged net barter terms of trade index	20.6%	188	42	1980-2021	4.60	0.29	2.93	6.77

Appendix Figure A1. Correlation matrix for model choices



B Appendix: Robustness checks

Given the very low threshold set for considering a choice “plausible”, one concern might be that we are adding several less relevant models to the TMS and thereby adding noise to the overall distribution and analysis. To assess sensitivity of our TMS approach and the benchmark results for democracy and growth, we re-constitute a more restrictive TMS after making several extra restrictions by excluding what we consider less plausible choices.

First, we omit specifications that do not include any control for time trends. Given the long time series we include, and the trended nature of both income and democracy, we anticipate substantial confounding for specifications omitting time-trend corrections.

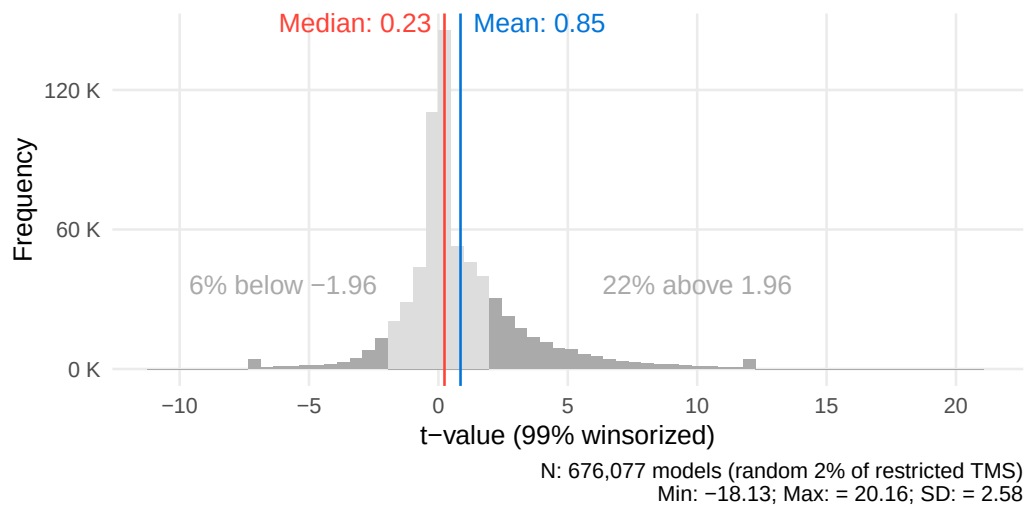
Second, we omit specifications that do not account for country-fixed effects. We believe there are very strong reasons to expect that our list of covariates (and especially specifications including only a few at the time) accounts for all relevant country-specific confounders that are fairly stable over time, for instance related to geography, political histories or cultural traits.

Third, we omit specifications using “classical” errors without any correction. Due to the time-series nature of our data and substantial cross-country heterogeneity, we do not believe that errors are independent and identically distributed (iid).

Fourth, we exclude specifications without any forward-lag on the DV, as we find it unlikely that any effect of regime type on growth would manifest in less than a year.

Fifth, we exclude the sample specification with developing countries only, as this throws out relevant information – especially in the long time series specifications, but also for countries such as Spain, Portugal and Greece in the short time series specification – from countries having undergone regime changes in currently developed countries, which are typically democratic at present. This is especially important if there are systematic selection biases for instance from economically successful countries having more stable regimes (perhaps especially democracies; Przeworski et al. 2000) and being more likely to exit the developing countries only sample.

Appendix Figure B1. Distribution of t-values, restricted TMS



Appendix Table B1. Results for random 2% of the restricted TMS

	OLS Coef.	XGBoost Imp. (%)
GDP data source (ref: Fariss et al.)		17.04
Maddison	-0.58	8.28
WDI	0.32	6.45
PWT	0.05	2.31
Imputed sample (ref: no)	1.20	15.79
Standard errors (ref: hetero)		12.80
Two-way	-0.62	7.22
Cluster	-0.63	5.57
Covariates (ref: life exp.)		11.30
Logged Terms Of Trade	-0.20	2.49
Inequality	-0.16	1.61
Economic Freedom	-0.30	1.25
Logged Population	-0.16	1.05
Logged Inflation	-0.04	0.93
Education	0.01	0.79
Logged Natural Resources	-0.19	0.69
Fertility	0.04	0.65
Logged Investment	-0.04	0.53
Logged Gov. Spending	-0.07	0.51
Logged Trade Openness	-0.06	0.42
Ethnolinguistic Frac.	0.03	0.11
Population Growth	-0.01	0.10
Internal Conflict	-0.01	0.09
International Conflict	0.02	0.09
Time-frame: Contemporary (ref: Historical)	-0.84	9.56
Democracy index source (ref: Polity V)		7.85
V-Dem Liberal Dem	-0.06	1.93
Lexical	-0.33	1.29
BMR	-0.44	1.18
Dd (Expanded)	-0.23	1.09
Polyarchy (From Observables)	-0.16	0.81
UDS	0.00	0.79
Polyarchy (V-Dem)	-0.23	0.77
Growth: first-dif (ref: logged)	-1.41	7.62
Democracy stock (ref: level)	-0.46	5.84
Lag structure (ref: 1 year)		5.53
10 Years	0.57	3.69
5 Years	0.35	1.84
Trend correction: Linear trend (ref: dummies)	0.59	4.33
Countries: Pop over 1 mil. (ref: all)	-0.13	1.21
Lags of GDP (ref: 1-year lag)		1.14
1-, 2-, And 3- Year Lag	-0.04	0.62
1- And 2- Year Lag	0.01	0.52
Out-of-sample RMSE	2.28	0.39
Adjusted R2	0.22	
Num. obs.	620,421	620,421

C Appendix: Further ideas for the future

- Discuss how features in the model space are a "compound treatment" — they can alter many things, including the sample composition (which is why we might want to focus on/compare all results against subset only using the imputed data, where we can hold sample size constant.)
- Mark our histograms of t-values with (a) previous influential results; (b) previous meta-analysis results.
- Introduce a GDP penalty for autocracies à la [Magee and Dozes \(2015\)](#).
- Generate a specification curve plot ([Simonsohn, Simmons, and Nelson, 2020](#))
- Construct a placebo test with simulated data ([Muñoz and Young, 2018](#)).
- Consider the possibility of bootstrapping. Instead of running each model within TMS across the full sample, one might adopt simulated samples. This has the added advantage of providing confidence intervals, which may be useful for the analysis of modeling choices.