Measuring State Capacity in the U.S. States

July 7, 2026

Abstract

Political scientists recognize the importance of state capacity for issues such as peace, economic development, public health, and regime stability. However, researchers have not agreed on how to measure state capacity. Crucially, many metrics inadvertently conflate strategic policy choices or policy liberalism with capacity. Furthermore, common cross-national measures do not work at subnational levels. American politics literature has largely overlooked subnational variations in state capacity. We propose a subnational measure of state capacity that compares all U.S. states from 2000 to 2019. Using a Bayesian factor analytic approach, we model state capacity as a reflective, latent trait. Our measure draws on shared policy goals to identify states' latent ability to deliver on policy, thus distinguishing the capacity to act from political preferences. We then run a series of validation tests to establish construct validity and to explore the political forces that have impacted subnational capacity in the American context.

Introduction

All government outputs are a combination of the choice to act and the capacity to convert those wishes into results. The ability to implement public policies that have been adopted is termed *state capacity* (Berwick & Christia 2018). State capacity has been linked to several important outcomes, from improvements in public health and human rights, to economic growth, to reductions in civil conflict (Acemoglu *et al.* 2015; Besley & Persson 2010; Englehart 2009; Evans & Rauch 1999; Fearon & Laitin 2003; Ziblatt 2008).

Political science has many tools to measure what states want to do. Party agendas and ideology, for instance, often describe specific policy objectives. Based on these metrics, states' specific policy preferences are well-traversed in the American and comparative literature. The *ability* to act on these goals, however, cannot be measured directly. As such, the substantial comparative literature on the effects of state capacity is not as firmly grounded empirically. This is particularly true for studies of state politics and policy within the United States, where such measures are not habitually employed in evaluating divergent outcomes among the states. In fact, there is currently no validated subnational measure of American state capacity.

Previous research has conceptualized and measured state capacity in various ways. The comparative literature has looked at this concept in terms of the relationships between state actors (e.g., bureaucratic efficiency) and in terms of specific policies (e.g., tax collection) and resources (e.g., wealth). These disparate measurement approaches limit our understanding of the overarching concept and its empirical relationship with other phenomena of interest. Furthermore, existing operationalizations of state capacity employed in the comparative literature, both cross-nationally and subnationally, are not suitable for the U.S. context. Our study, therefore, aims to improve construct validity, which concerns how well a measure operationalizes a theoretical concept.

These extant metrics are limited in crucial ways. First, they do not sufficiently distinguish the strategic choice to act from the ability to act. Setting tax rates, for instance, reflects political preferences as much as it reflects the capacity to collect taxes. Second, measures used in cross-national studies may not exist at subnational levels or may not be appropriate for the U.S. context. We must overcome both sets of difficulties to capture the variation of state capacity in the American states.

We introduce a new measurement of subnational American state capacity. In order to identify states' ability to enact their goals, we present a reflective theory of measurement that draws on 10 public policy goals shared by all states and develop a Bayesian factor analytic model built on these same goals. The states that can enact these shared goals demonstrate greater capacity. The policy items, encompassing goals in public health, education, crime, and economic well-being, are measured at the state level from 2000 through 2019, allowing us to identify subnational variation over time.

Next, we validate our measure using several convergent and divergent validity tests. We then explore the political forces that have impacted states' capacity levels. We find that liberal – but not necessarily Democratic party-controlled – states have developed higher levels of state capacity. On the other hand, states that have wasted resources, such as expending resources on redundant and segregated institutions in the Jim Crow period, have lower levels of state capacity decades later. State capacity can also be destroyed. We show, for instance, that Hurricane Katrina temporarily impeded state capacity in Louisiana. Thus, we show that state capacity is a political phenomenon apart from setting specific policy goals and the desire to act upon these goals. The conclusion discusses when this metric is applicable and when it should not be utilized. It also considers avenues for additional research on subnational variations in state capacity in the American states.

What is State Capacity?

State capacity is the ability of the executive branch to implement policies and perform core functions, including providing public goods, such as the maintenance of law and order, basic infrastructure, and administrative services. This definition is similar to Berwick & Christia (2018) and Hanson & Sigman (2021) and also relates to the power of the state to make citizens comply with laws and public policies they would not otherwise do (Lindvall & Teorell 2016; Soifer 2008).

Comparative literature has converged on three main theoretical dimensions of capacity but has not found evidence that they are empirically distinct. The extractive dimension relates to the ability of the state to collect resources and maintain information, such as taxes. This area has been referred to as the relationship between rulers and resource holders (Berwick & Christia 2018). A coercive dimension relates to the state's ability to secure compliance with public goals and depends on how well institutions structure state-agents and non-state agent interactions. The ability of the state to maintain and enforce law and order within its territory, for example, falls under this area. Third, states must coordinate collective action between state agents and society. This administrative dimension structures the relationship between politicians and bureaucrats, for instance, to deliver public services and regulate commercial activities. In practice, though, these dimensions are so interrelated and endogenous that scholarship has not found empirical evidence of separable dimensions (Hanson & Sigman 2021; Hendrix 2010).

Challenges to Measuring Subnational State Capacity

Although there is consensus that conceptually, state capacity refers to a state's ability to make and implement policies Berwick & Christia (2018), there is no commonly agreed upon measure of state capacity at the national level for a number of reasons. First, studies do not always separate capacity from the political desire to make and implement *specific* policy decisions. Comparative research has shown, however, that states must want to deploy resources effectively and desire to implement policies to achieve a given outcome (e.g., Harbers & Steele 2020; Holland 2016; Soifer 2008). Common indicators used to proxy state capacity, such as tax extraction and military spending, are not separable from political decisions and strategy since tax policy and military budgeting are political decisions (Andersen *et al.* 2014; Englehart 2009; Hanson & Sigman 2021). Not all countries want a large military or high tax rates, whether or not they *could* impose them, so such measures paint a misleading picture of their capacities. Related to this, studies do not always separate capacity from other concepts such as governance, liberalism, transparency, and the rule of law (Fukuyama 2013; Savoia & Sen 2015). This conflation impedes causal theorizing as well as the analysis of the relationship between capacity and other constructs.

Capacity may also vary within countries and across agencies as much as capacity varies between countries (e.g., Foa & Nemirovskaya 2016; Gingerich 2013; Harbers & Steele 2020). The challenges of producing replicable and reliable operationalizations of state capacity at the national level are compounded when measuring subnational capacity (Luna & Soifer 2017). At the subnational level, prior studies have adapted their concepts and measures to a particular country (often developing countries) or historical periods. Thus measures are usually not comparable across countries since indicators of capacity in one context may not exist, be unreliable, or be irrelevant for other countries. Notable subnational measures of capacity, such as colonial infrastructure (Acemoglu *et al.* 2015), age heaping in census reports (Lee & Zhang 2017), government repression of protest (Sullivan 2021), or meritocratic recruitment and predictability of bureaucratic careers (Bersch *et al.* 2017) are not appropriate for the U.S. states. Contemporary U.S. state institutions and structures were formed after independence; age heaping is not prevalent in developed countries; and protest repression is limited in consolidated democracies, as are non-meritocratic and unpredictable civil service careers. All U.S. states have more than minimal functionality, so developmental measures are insufficient.[1]

Survey-based measures of capacity, on the other hand, ask residents, bureaucrats, or experts about state institutions and services (e.g., Evans & Rauch 1999; Kyle & Resnick 2019; Luna & Soifer 2017). Public measures suffer from subjective biases and are infeasible. Widescale surveys require substantial capacity to conduct, and they are expensive to conduct regularly nationally, let alone to have sufficient sample size to identify subnational variation. Even large surveys like the ANES can have small samples or no responses in some U.S. states. While expert surveys are possible, given a sufficient number of cooperating state-specific experts (a la V-Dem), expert surveys would also require additional data collection and cost. In sum, the literature has, as yet, not produced a validated measure for state capacity across the American states.

Another crucial consideration is how the measurement of state capacity is formed, as existing measurements do not draw on a clear theory of measurement. In this regard, are measures based on the potential inputs into capacity (e.g., resources) or on the outcomes of capacity? For example, Lindvall & Teorell (2016) propose measuring state capacity according to the financial, human capital, and informational resources required to implement policies. By contrast, studies drawing on indicators of public goods provisioning, such as infant mortality or crime rates, are measuring the outputs of capacity. Several influential studies, in fact, mix inputs and outputs in forming their measurements (Bäck & Hadenius 2008; Fortin 2010; Hanson & Sigman 2021; Hendrix 2010; Soifer 2012). As will be discussed in the following sections, drawing on both inputs and outputs means mixing different types of measurements. Operationalizing a theoretical concept requires a coherent theory of measurement.

[1] We are not merely measuring "development" either. Besides the U.S. overall and all U.S. states scoring highly on the Human Development Index, our map differs from the HDI map. NY is one of the top HDI states, while NH is lower. State capacity, as we demonstrate later, shows the inverse. `globaldatalab.org`

Nonetheless, state capacity is an important concept in political science. Empirical research demonstrates that high capacity states provide more public goods and services than low capacity states. Higher capacity is associated with better health (Cingolani *et al.* 2015; Ziblatt 2008; Hanson 2015; Bustikova & Corduneanu-Huci 2017) and educational outcomes (Hanson 2015; Knutsen 2013; Lee & Zhang 2017). Capacity also reduces crime, corruption, and human rights violations (Englehart 2009; Cruz 2011; Richani 2010; Herrera & Martinelli 2013; Sullivan 2021). Low capacity states, on the other hand, are more prone to violent conflict and civil wars (Fearon & Laitin 2003; Besley & Persson 2010; Hendrix 2010). Thus, understanding variations across the U.S. states is important for understanding subnational disparities. While there are many things states may wish to do, high capacity states will be better able to accomplish those things.

Although we are examining the American case, subnational state capacity could be measured by the same technique in any country with subnational agents. Only a fully unitary state would be exempted from such analysis. The distinction will likely be strongest in fully-federalized systems, but any level of subnational control would be sufficient. Our focus, however, is on the U.S. states.

A Reflective Theory of Measurement

We introduce a new, reflective measurement of state capacity. Our construct is built on multiple outcomes that *reflect* states' ability to enact their will. The outcomes are separate from the particular policy preferences themselves.

Our theory of measurement requires us to differentiate between two types of measures: reflective and composite. A composite measure requires knowing all of the factors that make up a construct and including all of these components. A classic example of a composite measure is socio-economic status, defined as the combination of education, income, and occupational prestige (Baker 2014). If one part is not included, then the index would be incomplete and not measure socio-economic status (Stenner *et al.* 2008). While we have an understanding of what state capacity is – the ability of a government to enact its will and implement policies that it desires – its component parts are indeterminable and data are nonexistent for many components of capacity. Some researchers have attempted to use just one input component, the number of state employees, as a metric of state capacity. However, as noted, that is just one potential facet of state capacity, and it is thus incomplete. It also risks corruption from corruption – state governments may have extra employes *because* they are ineffecient or because they are using jobs for patronage. Thus, we could reasonably expect this metric to correlate with state capacity but we do not accept that it itself is state capaity.

Furthermore, even where the literature has claimed that sub-types of capacity exist, empirical studies do not find multiple dimensions or necessary components that map onto these theoretical sub-types (Hanson & Sigman 2021; Hendrix 2010), resulting in great ambiguity as to how components of capacity should be calculated and combined. Thus, the concept of state capacity is not theoretically or empirically amenable to a composite measurement strategy.

A reflective measure, on the other hand, is conducive for operationalizing state capacity. Reflective measures examine multiple outputs of a force and use *latent trait* modeling to identify this force in these results. The latent construct exists independently of the items used to measure the construct. For example, intelligence is a latent ability assessed through various types of tests. IQ tests take test question responses as reflections of an individual's underlying ability. In this case, there is no complete corpus of intelligence components to be assembled. As state capacity is in this mold, we employ the reflective measurement approach.

[Figure 1 about here.]

A diagrammatic version of this discussion is found in Figure **??**. Fundamentally, reflective and composite measures have different theories of causality: do the inputs make the construct or does the construct make the inputs? Reflective measures operate from the assumption that the variables used to measure the construct are *caused by* the underlying latent construct. Composite measures imply the opposite relationship; the construct only exists insofar as it is created by the various inputs. We argue that conceptually, state capacity works more like the former rather than the latter, given the inability to cognize all the features that combine to produce capacity.

As noted previously, some prior studies have combined inputs and outputs to operationalize state capacity, merging the two different measurement strategies. Mixing inputs and outputs functionally introduces circularities in the measurement, which cause difficulties in applying the measurement in causal tests. We avoid this problem by focusing on outputs in our reflective latent measure of state capacity.

Some scholars have expressed concerns about utilizing the outputs of capacity because they could be affected by exogenous factors, not just state capacity (Fukuyama 2013; Soifer 2008). The latent trait modeling strategy is crucial here. Using a Bayesian factor analytic model, we are isolating the element of state capacity *reflected* in these points from the exogenous influences impacting any particular outcome.[2] Another concern is that these measures risk becoming tautological (Brambor *et al.* 2020; Lindvall & Teorell 2016). This concern is not trivial, but it is not unique to measurements of state capacity. It is addressed by not applying the measure of state capacity to the outcomes that are used to form the measurement. For instance, studies that use infant mortality as a measure – a common technique – should not then apply that measurement in studies of childhood survival. Any scholar employing a measure – either composite or reflective – must be cognizant of this risk in utilizing measures or proxies. We are arguing that state capacity is the means by which states pursue or achieve these goals, so there is a causal implication; however, we are not testing the causal relationship between state capacity and these indicators in this study.

To form our reflective measure of state capacity, we must look for common goals among states. It cannot be an exhaustive list, because there are many things some states wish to do that others do not wish to accomplish. Those items would thus be polluted by political choice and strategy. For instance, a strong welfare state requires capacity, but not all administrations *want* to have a strong welfare state. State capacity is the ability to enact decisions, so we must separate political will as much as we can. We do this by looking only at outcomes for which all states could reasonably be assumed to share the same goal, even if they might disagree about the means of achieving that goal. For instance, all governments should oppose infant mortality (Bustikova & Corduneanu-Huci 2017; Ziblatt 2008) and violent crime (Englehart 2009; Richani 2010; Herrera & Martinelli 2013).[3] Individual politicians may have different hierarchical preferences among the items, so having a variety of indicators increases the accuracy of the scores. The included items will be discussed in the following section. Latent trait modeling identifies that latent ability – state capacity – that

[2]It is possible that exogenous influences could overwhelm a latent trait model for a particular indicator, though this would be readily apparent in the diagnostics of the model itself. We include numerous diagnostic tests to eliminate this as a possibility in Appendix 1.1. Furthermore, it should also be apparent theoretically and in the loading of that indicator.

[3]We want to be very clear that by removing political political choice and strategy we are not saying that we are removing political agency. The states certainly make choices to create capacity or not with investment over time. The objective (and, in fact, the challenge) is to avoid using metrics where the choice to act is seen more as a function of ideology or policy preferences than of the capacity to enact a wish that should be widely shared.

enabled the states to achieve, to varying degrees, their shared desires.

A New Measurement

We faced several challenges in producing a cross-temporal measurement of state capacity for the American states. First, we needed to identify outcomes that all the states would want to achieve in order to minimize the influence of political choice or strategy. Second, these had to be items measured at the state level. Third, they had to be measured with great frequency over time with the possibility of change over time.

Very few outcomes fit the first requirement. Many policy goals are not strongly desired by both political parties (Erikson *et al.* 1993; Grossman & Hopkins 2016; Grumbach 2018). This step is crucial. We minimize the impact of political will in this stage by removing as much as is practicable differences in the desire for an outcome, leaving us with a state's capacity to act. While parties may disagree about the best means of achieving literacy or preventing murder, those are shared objectives. The refinement by latent trait modeling helps additionally, but the true separation of political choice to act and policy liberalism happens at the indicator selection stage.

Data coverage across all states and for many years also eliminated some indicators. For instance, we were confident that the political leaders would favor lowering maternal mortality. However, data was not available for every state-year.[4] Thus, some theoretically potential items were excluded in practice.

The indicators we selected vary widely: the economy, health, education, and law enforcement. This ensures we are not measuring only one policy domain. We utilize the following items: infant mortality, poverty rate, robbery rate, car theft rate, murder rate, property crime rate, percent of the population with a BA or higher degree, percent of the population that dropped out of high school, elementary students with basic reading skills, and elementary students with basic math skills.[5] Several of these items have been used to proxy state capacity or as direct outcomes of state capacity in prior cross-national analyses:

- Infant mortality (Bustikova & Corduneanu-Huci 2017; Cingolani *et al.* 2015; Foa & Ne-

[4]The CDC suppresses the mortality rate data where the rate is sufficiently low.

[5]Most of these items are measured at the behest of the federal government – defined by Congress – and according to standards set by the federal government.

mirovskaya 2016; Hanson 2015).

- Crime (Foa & Nemirovskaya 2016; Englehart 2009; Cruz 2011; Richani 2010; Herrera & Martinelli 2013; Soifer 2012).

- Poverty (Acemoglu *et al.* 2012; Andersen *et al.* 2014; Fearon & Laitin 2003; Foa & Nemirovskaya 2016)[6]

- Primary or secondary education enrollment, literacy (Acemoglu *et al.* 2012; Hanson 2015; Lee & Zhang 2017).

Our position is not that these proxies for capacity were poor choices in the prior literature. Rather, we acknowledge that any single outcome of capacity would be a noisy measure of capacity. We do not argue that our indicators and differences among the states are *only* caused by state capacity. However, the latent trait modeling reduces the noise in the overall measurement by refining the state capacity from these several outcomes. Thus, the reflective measurement strategy is stronger than a single-item proxy operationalization of state capacity.

Certain items previously used in comparative work were excluded here. Some were excluded because they are inputs, such as the number of state employees. Others are excluded because, while linked to capacity, they are substantially functions of political choice in the US context. For instance, all US states *can* extract taxes, but the extent of the tax revenue is determined by political choices. Even the number of employees is a function of the political appetite for growing the state apparatus (or for awarding sinecures as patronage). Thus, the baseline level of state capacity in the US states means that some of these elements may be used in correlations to demonstrate validity, but they cannot measure state capacity at a subnational level in the United States.

We model a state's latent capacity using a Bayesian factor analytic method, described below. This approach is functionally very similar to Grumbach (2022, 2023), which is in turn very similar to Treier & Jackman (2008). Hanson & Sigman (2021) also use a latent trait model in the most comprehensive cross-national measure of state capacity over time. The measurement is dynamic, allowing each state's ideal point change per year.

[6] These studies use GDP as a proxy for poverty (except Acemoglu *et al.*; Fearon & Laitin). We use poverty rate directly. Although wealth and capacity are correlated – see Table 2 – they are not the same.

This latent variable analysis lets observed relationships between the capacity indicators determine how each indicator should affect states' capacity scores. Essentially, this strategy estimates an "ideal point" on a latent dimension for each state-year that best predicts the values of capacity indicators in the observed data. In particular, we use Bayesian factor analysis for mixed data because the capacity indicators are rates and counts.

The items and their discrimination parameters are shown in Table 1.[7] Discrimination parameters in this model are functionally analogous to factor loadings in frequentist analysis. Essentially, they represent the slope of the relationship between an indicator and a state's latent state capacity. Several variables were compiled from Grossmann *et al.* (2021), including the poverty rate, robbery rate, car theft rate, murder rate, and property crime rate. We extended the time span of many variables to cover years between 2000 and 2019 by compiling additional data from the original sources.[8] We used the "mice" package in R to perform multiple imputation on the variables, using a mixture of predictive mean matching, penalized regression, and Bayesian linear regression. We only had intermittent missingness patterns outside of the test scores, which were only recorded every other year – we interpolated the missing years for those scores.

[Table 1 about here.]

For a reflective measure, it is necessary to test for dimensionality. A scree plot shows that there is a strong first dimension (Appendix A1). We considered whether a meaningful second dimension exists; however, no other dimension has an eigenvalue above 1. As noted above, scholars have postulated multiple dimensions of capacity but have not identified them empirically. Not empirically idetifying a second dimension is consistent with what Hanson & Sigman (2021)'s findings. All items load well onto the first dimension, shown in Table 1. Next, Chronbach's Alpha is 0.83 and Guttman's 6 Lambda is 0.90 (Appendix A6), showing that the indicators are internally consistent. As such, we have strong reason to believe that a single latent dimension can be validly extracted. In the following section, we demonstrate that the extracted dimension is indeed state capacity.

According to our theory of measurement, the items that reflect state capacity should be well

[7] For identification purposes, we constrain 'Infant Mortality' to be negative and 'Basic Reading' to be positive. Our model is not sensitive to those variables in particular, though some restrictions are necessary given the Bayesian modeling. This was built out of the Bayesian factor analysis functions from the MCMCPack package in R. We run a 50,000 iteration Gibbs sampler with a burn-in of 1000.

[8] A complete list of the original data sources is available in the Appendix (A3 and A4).

correlated with one another and be reliable. If we assume that errors in the proxies are uncorrelated, then looking at more indicators for the latent trait will improve precision. In other words, the 10 indicators we include offer a clearer signal than using any one proxy for state capacity by itself. On the other hand, as a reflective measurement is focused on an underlying latent trait, the individual items themselves are not crucial to the metric (Coltman *et al.* 2008). We demonstrate that our measure of state capacity is robust to removing individual indicators. The scores generated in these leave-one-out models are nearly identical and are highly correlated with each other (r > 0.9 in all cases, see Appendix Figure 1). They strongly correlate with the metric used in this manuscript (r > 0.97).[9] In addition, Appendix A8 shows that the individual indicators are strongly correlated with the overall latent score (0.39 to 0.72), even when the item is dropped from the model (0.26 to 0.66). All of these diagnostics demonstrate that the individual indicators are reliable proxies and that the extracted scores tap into the latent, *reflected* dimension of state capacity.

We could also consider that these items from which we form our reflective measure, while universally desired, may not be equally desired in a systematic fashion. We have sought as much as possible to avoid this possibility when choosing items. However, to address this potentially, we examine the latent trait identified by comparing scales that could be construed as more targeted to Democrat or Republican policy priorities. For the "Democrat" measure, we drop the crime indicators. For the "Republican" measure, we drop the education metrics. We find that the resultant latent trait measures correlate with each other at r=0.5. Each correlates with our overall measure r=0.83 (Democrat) and r=0.7 (Republican). As several items are dropped in each of these cases, though, the measurements are not as strong in our opinion. This affirms our argument that we are measuring state capacity rather than partisanship, ideology, or political choice and strategy.

[Figure 2 about here.]

The latent trait score places states relative to each other each year. During the 20-year span included in our dataset, the scores range from -2.95 to 2.34. States' scores generally increase over time, which is a positive statement from a development perspective. A table with the mean state scores over the years and the state rank is shown in Appendix A5.

[9]The individual items are not perfect predictors of the latent trait, though they are all correlated with it. These indicators continue to be correlated with the scores generated by dropping those items (Appendix A2); we also look at the reliability with one item dropped (Appendix A7). This is a function of the latentness of the trait.

The ratings for 2019 for each state are shown in Figure 2. In general, the highest scoring states are in New England, followed by the Upper Midwest. The lowest scoring states are in the South. Potential causes for these regional disparities are discussed in Section 6. The highest capacity states are New Hampshire and Vermont. The lowest capacity states are Louisiana and Mississippi. The map demonstrates the face validity of our metric.

The map itself provides a demonstration that this measurement does not merely identify liberalism. After all, New Hampshire and Idaho are not more liberal than California or New York, but they score higher on the state capacity metric. Furthermore, as will be shown below, policy liberalism at the state level is a weak, if significant, predictor of state capacity development ($\beta = 0.135$). If we were merely measuring liberalism, then liberalism would be more strongly correlated with the metric. Given these findings, we turn to additional empirical validation tests for our measure of state capacity.

Validation: Convergent and Divergent

This section examines the construct validity of our metric. Construct validity concerns how well the measurement translates the core idea into numbers. It uses a "relationalist" approach (Trochim & Donnelly 2001), which means looking for positive or negative (or null) relationships with other constructs that should be theoretically related. Our argument here is not that these items are causally related to state capacity. Rather, the items used to examine convergent validity can be thought of as alternative proxies for capacity. We can draw, for instance, on input measures or items used in other state capacity studies. For divergent validity, we choose items that are either not expected to be related to state capacity or are expected to be negatively related.[10]

Identifying "evidence for both convergent and discriminant validity" is considered "by definition [...] evidence for construct validity" (Trochim & Donnelly 2001). In other words, these validity tests tell us whether the measure of capacity is tapping into what we intend to identify by showing that the measure does relate to items to which we would theoretically expect our target construct to relate (convergent validity). Tests of divergent validity also indicate that we are able to discriminate between unrelated constructs and our target construct.

[10] All data sources for variables used in the validity tests are listed in Appendix A4.

To evaluate the significance of the relationship between each correlate and state capacity, we run linear regressions with state-and-year fixed effects. This approach ensures that the coefficients represent how much a unit change in the correlates impacts (or is impacted by) a unit change in state capacity *within the state*. Different baseline levels of any of the correlates will not therefore affect our estimates. The coefficient estimates of these regressions are shown in Table 2.

[Table 2 about here.]

Convergent Validity

Convergent validity demonstrates that "the operationalization is similar to (converges on) other operationalizations that it theoretically should be similar to" (Trochim & Donnelly 2001). To examine this type of validity for state capacity, we look to factors that should correlate positively with state capacity. First, states with a higher number of state employees should have higher capacity. As noted, this measure has been suggested in other cases and should be positively correlated with capacity even if it is a noisy signal. We measure public employees in two ways. We first look at the total number of employees working in government and government enterprises. We find a positive and significant relationship with state capacity. We also consider the share of the total workforce that is employed by the state government. This may be a weaker indicator of state capacity – having a larger share of the population employed by the state could happen without the state being more effective – but the relationship is positively correlated with state capacity.

Another common metric used to measure state capacity is state revenue from taxes. Notwithstanding the fact that tax policies are political decisions, we expect that states that collect higher tax revenues would on average have greater capacity. They have more money to spend on building capacity. We find a positive and significant relationship between state capacity and tax revenue (in billions USD), as well as between the share of the state's economy that taxes make up ($p=0.056$). Convergence with these statistics supports the premise that we are measuring states' latent capacity.

Divergent Validity

We also consider the discriminant or divergent validity of our measure. Divergent validity is the "degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should be not be similar to”(Trochim & Donnelly 2001). This can mean items with which it should be uncorrelated or items with which it should be negatively correlated.

Namely, we are allowing for the measurement to be uncorrelated with state policies that are mostly a function of political choices and ideology. We first look at economic inequality and find no relationship between state capacity and state-level income inequality. This lack of relationship makes sense, since the level of inequality within states does not clearly translate into aggregate capacity. Next, we expect a negative relationship between states' debts and their capacity. This is the other side of the tax collection efficacy above. States that lack financial resources are more constrained in pursuing their policy goals. They are also not demonstrating the managerial ability to raise taxes to match expenditures or keep expenditures limited to revenues. There is a negative relationship between the state's debt and capacity. Finally, there is a negative but insignificant relationship between the state's budget ratio and capacity. The budget ratio is how far off the state's income and expenditures are from each other. Diverging from these outcomes is consistent with tapping into states' latent capacity.

Validation: Plausibly Exogenous Shocks

State capacity is a latent ability that states develop over time. However, when a crisis strikes, states go to war with the army they have. In the case of Covid-19, for example, states that needed time to ramp up their ability were back-footed by an imminent crisis (Auerbach *et al.* 2024). Shocks can also hit the system itself, undercutting or destroying the capacity that states have built. An important test for the validity of our measure of state capacity will be to analyze how a particularly strong exogenous shock affects a given state's observed capacity. For this, we look at a massive natural disaster that wreaked substantial but localized harm in the US in this period – Hurricane Katrina. This is hardly the first paper to treat Hurricane Katrina as an exogenous shock and explore long-run differences caused by the hurricane, but this is the first to look at how it impacted state capacity (Malhotra & Kuo 2008; Deryugina *et al.* 2018; Schüwer *et al.* 2019; Rayamajhee *et al.* n.d.).

Hurricane Katrina, a category five hurricane, struck the Gulf Coast in 2005. It killed nearly 2000 people and caused 125 billion dollars (2005 USD) in damage – and almost all of the deaths were limited to Louisiana. Buildings were destroyed, people were displaced, school systems were shuttered, and hospitals were overwhelmed. We evaluate the effect of this hurricane using a difference-in-differences design to assess the effect of the exogenous shock. We code Louisiana as the "treated" state. The goal is to examine the relationship, as tightly as possible, to that one event. We examine the effects through 2010 since considerable state, federal, and NGO resources were invoked to resolve the damage rapidly.

[Table 3 about here.]

For this analysis, we employ a difference-in-differences (DiD) design.[11] We focus on 2000-2010, assuming that the effect of Katrina should diminish over time. 2005 and before are coded as pre-treatment, as the hurricane happened at the end of August 2005, and years 2006 and onward are counted as treated. For our estimation of the average treatment effect of the treated (ATT), we use the doubly robust DiD as our preferred estimator. We do so to reduce the possible effects of our covariates on differential pre-treatment trends – and recent literature has suggested that the doubly robust DiD is the most efficient way to do that (Sant'Anna & Zhao 2020; Roth 2022).[12]

We find in both the regression-adjusted difference-in-differences model and in the doubly-robust difference-in-differences model that there was a negative effect of Hurricane Katrina on state capacity in Louisiana (Table 3). In fact, the effect is more than doubled in the doubly-robust model.[13] Importantly, we would not expect the effects to endure. The governments responded to the disaster with programs and resources to fix the damage. We would expect such efforts to be effective, ameliorating a temporary blow to capacity. This is evidence of our theory that capacity can be damaged by disaster but that it is recoverable with work and resources.

[11] We considered a synthetic controls design, as we have only one treated case. However, Louisiana is the lowest-scoring state even pre-Katrina, so the synthetic state did not quite match in the pre-period. Thus, although the synthetic control model also showed a negative effect on Louisiana's capacity – similar to the DiD estimates – we do not include the results here. While this same set of concerns might be troubling to a DiD design, we include an event study plot in the appendix that shows we have parallel trends in the pre-treatment period (Appendix A3). The pre-period effects are also not statistically significant, another test for parallel pre-trends, though power concerns are a limitation here (Roth 2022).

[12] We did not use any of the other modern DiD specifications because we have a single treatment that is, by definition, not staggered.

[13] The effect is not significant if other states are included. This is consistent with the concentration of death and destruction in Louisiana, although other states were not entirely unaffected, during that event.

Predictors of State Capacity

We also use regression analysis to model several factors that affect the development of state capacity. Recall, state capacity is something states choose to build. It is then brought to bear on a particular need. If needs are known ex ante, then states can build capacity to serve ends; for instance, we expect states would have built capacity in order to fight crime. However, if a need is sudden and un-predicted, it cannot be rapidly built just to address that problem, such as an invasion or epidemic.

We examine several covariates that feasibly correlate with the development of state capacity. Some of these have been explored in cross-national literature. We expect a positive relationship with state wealth, since states with resources can use those resources, if they choose, to build capacity (e.g., Lindvall & Teorell 2016; Besley & Persson 2009, 2010). We use Gross State Product per capita to measure states' wealth.

We also consider demographic features. Research has shown that investment in the community and public goods goes down as populations diversify, including in the U.S. (Habyarimana *et al.* 2007; Suryanarayan & White 2021). To address race-based effects, we include population fraction non-white in the state on a 0 to 1 scale. We expect a negative relationship. Note, we are not making a claim about race groups but rather about the impact of public decision-making related to racial diversity (namely majority groups not wanting to develop resources for minority groups).

Social capital could have a positive effect on state capacity. It enables groups to band together and push the state to act (Brady *et al.* 1995). Thus, social capital could be turned to encouraging state capacity. However, the networks that build social capital could also cut out the state and make citizens less concerned with whether or not the state builds capacity. Subnational frontier regions also tend to have lower capacity but higher levels of volunteerism as they prefer a more limited role of the state (Foa & Nemirovskaya 2016). We thus could see either relationship. The measure for social capital comes from Hawes *et al.* (2013). To capture the frontier history of a state, we include longitude and latitude in Appendix A11.

Political forces are considered in two ways. One is a binary variable for Democratic governors. The other is the state policy liberalism score (Caughey & Warshaw 2016). Democrats are often accused of attempting to expand the state and its role, which may require building capacity. A liberal agenda, regardless of who is signing the bills, would have this effect as well.

State capacity is an investment states make over time. However, instead of making that investment, states might waste resources rather than invest in capacity. In American history, racial segregation is an major instance in which states were intentionally inefficient with resources. As such, it can provide a demonstration of that proposition. However, the underlying premise is not specific to race-based waste. For decades, many American states invested in segregation and other manifestations of racism, such as building separate (and unequal) facilities and enforcing barriers. These were resources that could not be spent on building capacity. More historically, some states made choices not to build state capacity as well. For instance, during the post-Reconstruction era in the American South, Suryanarayan & White (2021) show that, in areas where slavery was more prevalent, wealthy whites aligned with poor whites to build a race-based coalition that weakened bureaucratic capacity in order to prevent redistribution to former slaves. We propose that past mis-allocation continues to show in lower state capacity in those states. We assess this by including a binary indicator for states wherein the state or counties were covered under the Voting Rights Act (VRA).[14] We expect a negative relationship.

[Table 4 about here.]

It has also been suggested that a professionalized legislature would increase states' capacity. Squire (2007) explains the professionalism is the expertise and formalization of the government position: "Professionalism is typically associated with unlimited legislative sessions, superior staff resources, and sufficient pay to allow members to pursue legislative service as their vocation"(211). With these resources, legislators could build a more efficacious state. Individual legislators would have experience and know-how that they would lead to institutional development. We examine this using the legislature length, expenditures, and salaries variables from Bowen & Greene (2014).[15] We do not find a significant relationship between professionalism and capacity; those results are not shown. This could reflect measurement issues in the legislative professionalism metrics, or it could mean that professionalization – making the job full time or hiring more staffers – does not

[14] The whole state list was Alabama, Alaska, Arizona, Georgia, Louisiana, Mississippi, South Carolina, Texas, and Virginia. The additional states are California, Florida, New York, North Carolina, and South Dakota. We also ran models with only states where the full state was subject to the VRA; the relationship is robust and presented in Appendix A9.

[15] The variables were logged due to heavy skew.

mean that legislators next choose to develop the government capacity.

We use multilevel OLS models with the point estimate with random intercepts fit on year and state in the panel-type dataset. The models are necessarily parsimonious because of the low N – there are only 50 states. Table 4 shows our results. Though we can include few variables, we see strong explanatory power. For robustness, we also ran completely Bayesian models that incorporated the uncertainty of the latent trait measure in Stan, and the results replicated. For simplicity, the point-estimate models are shown here.

We find that larger states have higher capacity[16].

The wealth effect is not significant. This makes sense given that wealth is not capacity by itself; states would have to use wealth to build capacity. It also explains why economically impactful states like California do not score the highest; while they have economic power or can create policies that will make private companies want to fit that market, it does not mean the state can simply achieve all its goals.

Interestingly, we see that policy liberalism is associated with increased capacity; the party in control of the governorship, however, is not. The driving force is therefore the policy itself, rather than the party pushing it. This is theoretically validating, since the policy would be what drives the capacity, and Democratic governorship would be a loose proxy for liberalism.[17]

Consistent with our expectations, we find a negative relationship between the non-white population and state capacity. This fits with the comparative literature. We also find a positive effect of social capital on state capacity development. Our largest effect is shown in our VRA indicator. States that spent decades investing in racism instead of state capacity have worse capacity to this day. Notably, the effect is robust to the inclusion of the race variable, so it is not just measuring diversity. The effect size is substantial. A 1.4 unit change in state capacity is a shift from Mississippi to New York in 2019, holding all else constant. Substantial concerted effort over time would likely be necessary to correct these deficiencies.

[16]Places with growing populations may be attracting them through their efficacy. Population growth can encourage economic development and increase the tax base. Eventually these can also assist development. State capacity is not just population.

[17]An alternative measure of partisanship is shown in Appendix A10; it is also not significant.

Discussion and Conclusions

In this study, we have tackled the methodological problem of measuring state capacity sub-nationally. Although the comparative literature has found that state capacity drives crucial developmental and political actions, scholars have not settled on one metric for it. Furthermore, many metrics cannot report variation across subnational units. In this project, we have validated the first comprehensive measurement of state capacity for the 50 U.S. states. We provide 20 years of state-by-state state capacity scores. The panel data can be used to examine both causes and consequences of state capacity and differences therein.

We present a *reflective* measure of state capacity, extracting states' latent abilities to enact their will from their success in achieving shared goals. Notably, the focus on shared goals allows us to avoid conflating the political decision to act from the ability to act. We validate this measure with convergent and divergent validity tests. Our measurement is positively related to several factors the literature has used to proxy for state capacity. However, our measurement improves on these proxies by reducing the noisy measurement of single indicators.

We then turn to the correlates of state capacity. Partisanship is not a strong predictor of capacity, but policy liberalism is. State capacity is not a question of having Democrats or Republicans in office. It is about public investment.

An important manifestation of such investments is shown in cases where states have made the exact opposite choice. They have used resources on particular, rather than public, goals, at the expense of investments in the state or for the total population, which has undermined state capacity. We demonstrate this by showing the continued lower capacity in states covered under the Voting Rights Act. We are not arguing that racism itself undermines state capacity, that federal intervention against racism builds capacity, or that this is a race-specific phenomenon. We are also not arguing that states are inherently doomed because they cannot change the past. We are demonstrating that misspending resources undermines states.

Similarly, we are not arguing that unrestrained government spending causes state capacity. The point is that resources must be used *efficiently* if those resources are to build capacity. States must gather information on the populace in order to target resources effectively, which will require some level of state capacity. They must then choose to deploy those resources where they are most needed. In some cases, doing so would require political will and, occasionally, political courage. The state does not achieve its objectives by collecting money and/or writing large checks; it must put that money towards effective deployments of programs and personnel to achieve state objectives. The focus should therefore be on state investments rather than state expenditures.

The reflective measurement strategy offers many avenues for future research. What other factors cause states to develop their state capacity? What other forces explain national and subnational variations in capacity? Research could also probe the effects of state capacity on multiple domains and the potential for heterogeneous treatment effects. We particularly endorse the measure for understanding variation in policy outcomes and responses to exogenous shocks. This metric is also useful to Americanists engaged in between-state comparisons. Researchers should account for disparities in capacity in selecting their most-similar and/or most-different case designs. Capacity is otherwise a potential omitted variable.

Beyond the American context, our measure can also aid future research examining variations in the capacities of subnational governments in other countries. Researchers might consider Germany, Brazil, or India, as federalized systems, or the United Kingdom, which has devolved powers. More centralized states, like France, Japan, or South Africa, could also be considered. Naturally, in turn, the subnational capacity would not be pertinent to contexts in which no subnational authority or devolved power exists.

In regard to cross-national comparisons of state capacity, however, researchers should be cognizant of the differences in how states and their subnational governments measure individual items. For instance, many cross-national studies use items like infant mortality. However, if these measurements were done according to different standards among the countries, then the measurement would be corrupted. Additionally, any case where researchers suspect countries or subnational units in those countries of intentional deception cannot be readily compared. By using regulated items within the United States, we mitigate this concern. However, cross-national comparisons should be cognizant of this point.

We reiterate such caution in the application of this measurement. No measurement tool, whether composite or reflective, should be used to measure the items that went into creating it. To do so introduces circularity into the modeling strategy. Thus, in our case, we would advise against using the 2019 capacity scores to predict 2019 homicide rates variation across the states. Scholars interested in employing our measure are thus best served by being aware of what is under the hood. To that end, we have provided those details in this study.

Similarly, scholars may be considering a complementary or compounding relationship between state capacity and public capacity. For instance, private groups could also be acting on a problem or acting in concert with the state through public-private partnerships, which require both public and private capacity to form and operate (Berwick & Christia 2018). In that case, the modeling strategy could incorporate covariates like state-level social capital metrics. Capacity outside the state is distinct theoretically and empirically from state capacity.

Finally, state capacity is sticky over time. A country or subnational unit's capacity does not suddenly spring into or fall out of existence. Thus, researchers interested in using our data across multiple years should account for that temporal linkage. We nonetheless encourage scholars to take advantage of the two-decade time span in our model to select years most proximal to their policy targets.

Democratic representation assumes elected officials will enact policies their citizens desire. Without capacity, though, states merely converts preferences into legislation, not results. Thus, the ability of American states to enact their will has a profound influence on their ability to translate democratic outcomes into lived experiences. With this measurement, scholars can now investigate the impact of that capacity on states' fulfillment of the democratic mission.

References

Acemoglu, Daron, Johnson, Simon, & Robinson, James A. 2012. The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, **102**(6), 3077–3110.

Acemoglu, Daron, García-Jimeno, Camilo, & Robinson, James A. 2015. State capacity and economic development: A network approach. *American Economic Review*, **105**(8), 2364–2409.

Andersen, David, Møller, Jørgen, Rørbæk, Lasse Lykke, & Skaaning, Svend-Erik. 2014. State capacity and political regime stability. *Democratization*, **21**(7), 1305–1325.

Auerbach, Kiran, Lerner, Joshua Y., & Ridge, Hannah M. 2024. State Capacity and Covid-19 Responses: Comparing the U.S. States. *State Politics & Policy Quarterly*, 1–21.

Bäck, Hanna, & Hadenius, Axel. 2008. Democracy and state capacity: exploring a J-shaped relationship. *Governance*, **21**(1), 1–24.

Baker, Elizabeth H. 2014. Socioeconomic status, definition. *The Wiley Blackwell encyclopedia of health, illness, behavior, and society*, 2210–2214.

Bersch, Katherine, Praça, Sérgio, & Taylor, Matthew M. 2017. State capacity, bureaucratic politicization, and corruption in the Brazilian state. *Governance*, **30**(1), 105–124.

Berwick, Elissa, & Christia, Fotini. 2018. State capacity redux: Integrating classical and experimental contributions to an enduring debate. *Annual Review of Political Science*, **21**, 71–91.

Besley, Timothy, & Persson, Torsten. 2009. The origins of state capacity: Property rights, taxation, and politics. *American economic review*, **99**(4), 1218–44.

Besley, Timothy, & Persson, Torsten. 2010. State capacity, conflict, and development. *Econometrica*, **78**(1), 1–34.

Bowen, Daniel C, & Greene, Zachary. 2014. Should we measure professionalism with an index? A note on theory and practice in state legislative professionalism research. *State Politics & Policy Quarterly*, **14**(3), 277–296.

Brady, Henry E., Verba, Sidney, & Schlozman, Kay Lehman. 1995. Beyond SES: A Resource Model of Political Participation. *American Political Science Review*, **89**(2), 271–294.

Brambor, Thomas, Goenaga, Agustín, Lindvall, Johannes, & Teorell, Jan. 2020. The lay of the land: Information capacity and the modern state. *Comparative Political Studies*, **53**(2), 175–213.

Bustikova, Lenka, & Corduneanu-Huci, Cristina. 2017. Patronage, trust, and state capacity: The historical trajectories of clientelism. *World Politics*, **69**(2), 277–326.

Caughey, Devin, & Warshaw, Christopher. 2016. The dynamics of state policy liberalism, 1936–2014. *American Journal of Political Science*, **60**(4), 899–913.

Cingolani, Luciana, Thomsson, Kaj, & De Crombrugghe, Denis. 2015. Minding Weber more than ever? The impacts of state capacity and bureaucratic autonomy on development goals. *World Development*, **72**, 191–207.

Coltman, Tim, Devinney, Timothy M., Midgley, David F., & Venaik, Sunil. 2008. Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, **61**(12), 1250–1262.

Cruz, José Miguel. 2011. Criminal violence and democratization in Central America: The survival of the violent state. *Latin American politics and society*, **53**(4), 1–33.

Deryugina, Tatyana, Kawano, Laura, & Levitt, Steven. 2018. The economic impact of Hurricane Katrina on its victims: Evidence from individual tax returns. *American Economic Journal: Applied Economics*, **10**(2), 202–233.

Englehart, Neil A. 2009. State capacity, state failure, and human rights. *Journal of Peace Research*, **46**(2), 163–180.

Erikson, Robert S, Wright, Gerald C, & McIver, John P. 1993. *Statehouse democracy: Public opinion and policy in the American states*. Cambridge University Press.

Evans, Peter, & Rauch, James E. 1999. Bureaucracy and growth: A cross-national analysis of the effects of "Weberian" state structures on economic growth. *American sociological review*, 748–765.

Fearon, James D, & Laitin, David D. 2003. Ethnicity, insurgency, and civil war. *American political science review*, **97**(1), 75–90.

Foa, Roberto Stefan, & Nemirovskaya, Anna. 2016. How state capacity varies within frontier states: A multicountry subnational analysis. *Governance*, **29**(3), 411–432.

Fortin, Jessica. 2010. A tool to evaluate state capacity in post-communist countries, 1989–2006. *European journal of political research*, **49**(5), 654–686.

Fukuyama, Francis. 2013. What is governance? *Governance*, **26**(3), 347–368.

Gingerich, Daniel W. 2013. Governance indicators and the level of analysis problem: empirical findings from South America. *British Journal of Political Science*, **43**(3), 505–540.

Grossman, Matt, & Hopkins, David A. 2016. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. Oxford University Press.

Grossmann, Matt, Jordan, Marty P, & McCrain, Joshua. 2021. The correlates of state policy and the structure of state panel data. *State Politics & Policy Quarterly*, **21**(4), 430–450.

Grumbach, Jacob M. 2018. From Backwaters to Major Policymakers: Policy Polarization in the States, 1970–2014. *Perspectives on Politics*, **16**(2), 416–435.

Grumbach, Jacob M. 2022. *Laboratories against democracy: How national parties transformed state politics*. Vol. 182. Princeton University Press.

Grumbach, Jacob M. 2023. Laboratories of democratic backsliding. *American Political Science Review*, **117**(3), 967–984.

Habyarimana, James, Humphreys, Macartan, Posner, Daniel N, & Weinstein, Jeremy M. 2007. Why does ethnic diversity undermine public goods provision? *American political science review*, **101**(4), 709–725.

Hanson, Jonathan K. 2015. Democracy and state capacity: complements or substitutes? *Studies in comparative international development*, **50**(3), 304–330.

Hanson, Jonathan K, & Sigman, Rachel. 2021. Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research. *Journal of Politics*, **83**(4).

Harbers, Imke, & Steele, Abbey. 2020. Subnational Variation Across States: A Typology and Research Agenda. *Latin American Politics and Society*, **62**(3), 1–18.

Hawes, Daniel, Rocha, Rene, & Meier, Kenneth. 2013. Social capital in the 50 states: Measuring state-level social capital, 1986–2004. *State Politics & Policy Quarterly*, **13**(1), 121–138.

Heckman, James J, Ichimura, Hidehiko, & Todd, Petra E. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, **64**(4), 605–654.

Hendrix, Cullen S. 2010. Measuring state capacity: Theoretical and empirical implications for the study of civil conflict. *Journal of peace research*, **47**(3), 273–285.

Herrera, Helios, & Martinelli, César. 2013. Oligarchy, democracy, and state capacity. *Economic Theory*, **52**(1), 165–186.

Holland, Alisha C. 2016. Forbearance. *American political science review*, **110**(2), 232–246.

Knutsen, Carl Henrik. 2013. Democracy, State Capacity, and Economic Growth. *World Development*, **43**, 1–18.

Kyle, Jordan, & Resnick, Danielle. 2019. Delivering more with less: Subnational service provision in low capacity states. *Studies in Comparative International Development*, **54**(1), 133–163.

Lee, Melissa M, & Zhang, Nan. 2017. Legibility and the informational foundations of state capacity. *The Journal of Politics*, **79**(1), 118–132.

Lindvall, Johannes, & Teorell, Jan. 2016. *State capacity as power: a conceptual framework*. Department of Political Science, Lund University.

Luna, Juan Pablo, & Soifer, Hillel David. 2017. Capturing sub-national variation in state capacity: A survey-based approach. *American Behavioral Scientist*, **61**(8), 887–907.

Malhotra, Neil, & Kuo, Alexander G. 2008. Attributing blame: The public's response to Hurricane Katrina. *The Journal of Politics*, **70**(1), 120–135.

Rayamajhee, Veeshan, March, Raymond J, & Clark, Corbin CT. Shock me like a Hurricane: how Hurricane Katrina changed Louisiana's formal and informal institutions. *Journal of Institutional Economics*, 1–20.

Richani, Nazih. 2010. State Capacity in Postconflict Settings: Explaining Criminal Violence in El Salvador and Guatemala. *Civil Wars*, **12**(4), 431–455.

Roth, Jonathan. 2022. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, **4**(3), 305–322.

Sant'Anna, Pedro HC, & Zhao, Jun. 2020. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, **219**(1), 101–122.

Savoia, Antonio, & Sen, Kunal. 2015. Measurement, evolution, determinants, and consequences of state capacity: a review of recent research. *Journal of Economic Surveys*, **29**(3), 441–458.

Schüwer, Ulrich, Lambert, Claudia, & Noth, Felix. 2019. How do banks react to catastrophic events? Evidence from Hurricane Katrina. *Review of Finance*, **23**(1), 75–116.

Soifer, Hillel. 2008. State infrastructural power: Approaches to conceptualization and measurement. *Studies in Comparative International Development*, **43**(3), 231–251.

Soifer, Hillel David. 2012. Measuring state capacity in contemporary Latin America. *Revista de Ciencia Política*, **32**(3), 585–598.

Squire, Peverill. 2007. Measuring state legislative professionalism: The squire index revisited. *State Politics & Policy Quarterly*, **7**(2), 211–227.

Stenner, AJ, Burdick, DS, & Stone, MH. 2008. Formative and reflective models: Can a Rasch analysis tell the difference. *Rasch measurement transactions*, **22**(1), 1152–1153.

Sullivan, Heather. 2021. Subnational state capacity and the repression of protest: Evidence from mexico. *Political Research Quarterly*, **74**(3), 587–598.

Suryanarayan, Pavithra, & White, Steven. 2021. Slavery, reconstruction, and bureaucratic capacity in the American south. *American Political Science Review*, **115**(2), 568–584.

Treier, Shawn, & Jackman, Simon. 2008. Democracy as a latent variable. *American Journal of Political Science*, **52**(1), 201–217.

Trochim, William MK, & Donnelly, James P. 2001. *Research methods knowledge base*. Vol. 2. Atomic Dog Pub. Macmillan Publishing Company, New York.

Ziblatt, Daniel. 2008. Why some cities provide more public goods than others: A subnational comparison of the provision of public goods in German cities in 1912. *Studies in Comparative International Development*, **43**(3), 273–289.

List of Figures

Figure 1: Reflective vs. Composite Measures

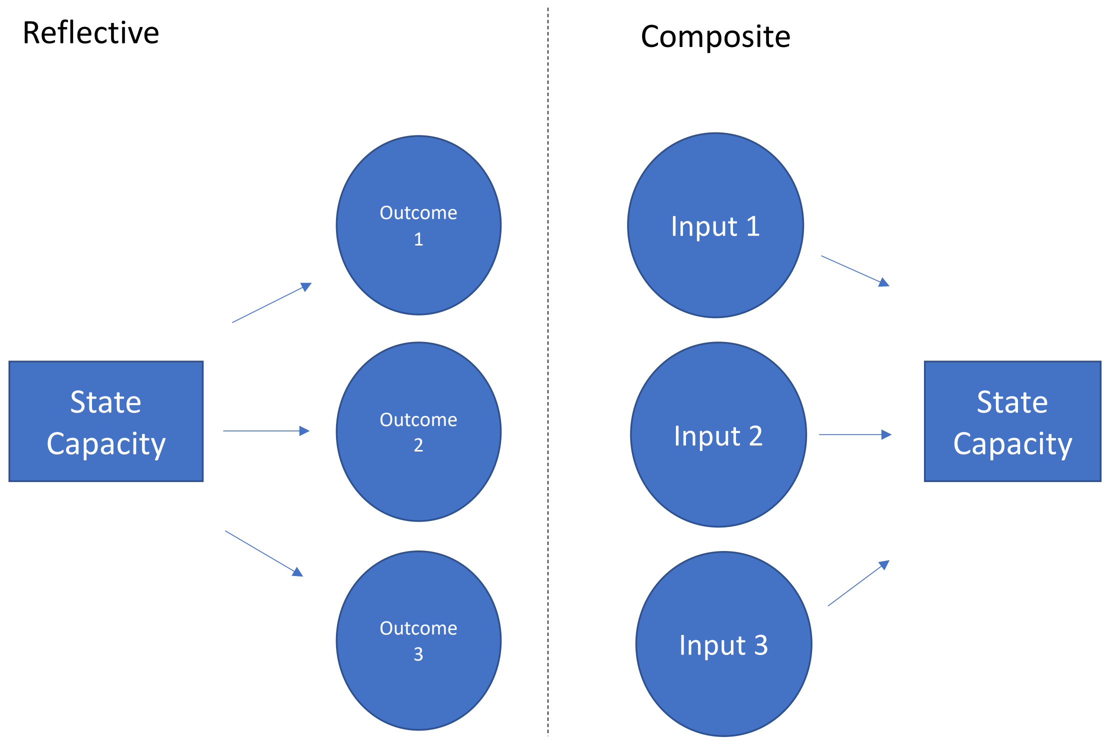
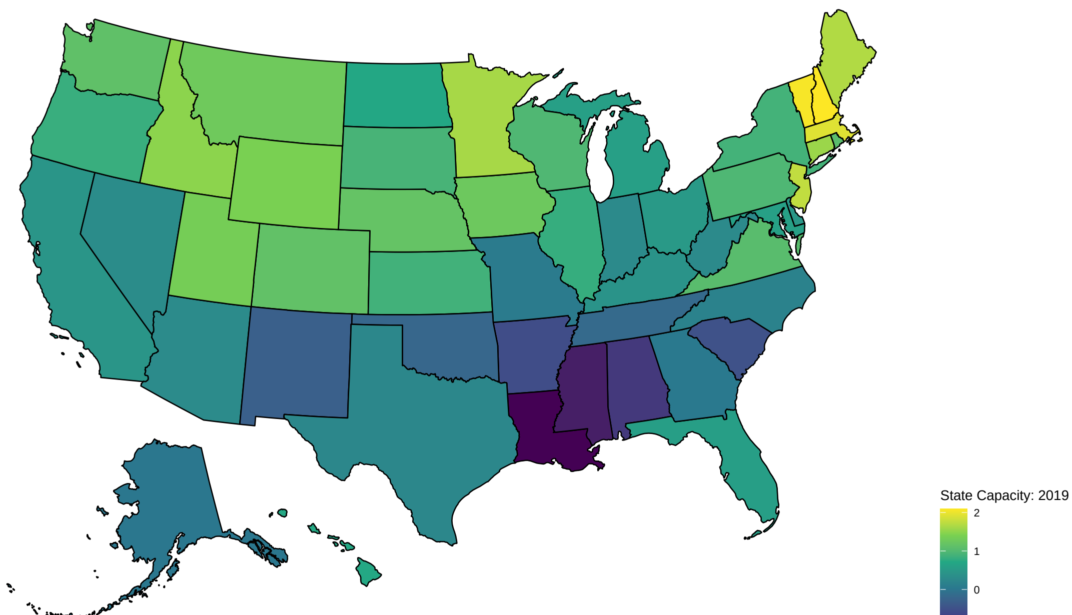


Figure 2: State Capacity in 2019



List of Tables

Table 1: Discrimination Parameter for State Capacity Factor

	Mean	SD
Infant mortality	-0.71	0.03
Poverty rate	-0.53	0.03
Robbery rate	-0.34	0.03
Car theft rate	-0.50	0.03
Murder rate	-0.70	0.03
Property crime rate	-0.69	0.03
BA degree	0.60	0.03
High school dropout	-0.70	0.03
Basic math	0.53	0.03
Basic reading	0.61	0.03

Table 2: Summary of Convergent and Divergent Validity Tests

Correlate	Coefficient	SE	p-value
Convergent Validity			
State Employees (in millions)	1.12	0.33	0.001
State Workforce	2.80	1.54	0.070
Population (in millions)	0.15	0.01	0.000
Tax Revenue (in billions $)	3.70	0.49	0.000
Divergent Validity			
Economic Inequality	0.40	0.32	0.217
State Debt (percent of GSP)	-0.02	0.01	0.022
State Budget Ratio	-0.04	0.03	0.247

Table 3: Difference-in-Differences on Effect of Hurricane Katrina on State Capacity

	DID 1	DID 2
ATT	-0.258^{*}	-0.947^{***}
	(0.130)	(0.238)
Num.Obs.	200	200
Model Specification	Regression Adjustment	Doubly Robust DiD
Years	2000:2010	2000:2010

*p<0.05, **p<0.01, ***p<0.001

DID 1 uses an aggregated event study DID design with a regression adjustment following Heckman *et al.* (1997). DID 2 follows the same specification but uses the Doubly Robust DID model of Sant'Anna & Zhao (2020). Our covariates are population, percent non-white, GSP per capita, and a dummy for if the Governor is a Democrat. Our treatment is for states affected by Hurricane Katrina.

Table 4: Mixed Effects Models on Correlates of State Capacity: Bayesian Measure

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−0.319	−0.207	−0.166	−0.190
	(0.214)	(0.207)	(0.205)	(0.208)
VRA	−1.613***	−1.457***	−1.229***	−1.079***
	(0.332)	(0.310)	(0.295)	(0.274)
Dem. Governor	0.004	0.004	−0.017	−0.014
	(0.015)	(0.015)	(0.016)	(0.019)
GSP per cap (thousands)	0.002	0.002	−0.001	0.002
	(0.002)	(0.002)	(0.002)	(0.003)
Population (millions)	0.113***	0.098***	0.071***	0.009
	(0.011)	(0.012)	(0.014)	(0.016)
Percent Non-White		−0.400*		
		(0.183)		
Policy Liberalism			0.135***	
			(0.039)	
Social Capital				0.023+
				(0.013)
Num.Obs.	1000	1000	750	480
R2 Marg.	0.343	0.326	0.311	0.247
R2 Cond.	0.981	0.977	0.978	0.974

+p<0.1, *p<0.05, **p<0.01, ***p<0.001