Tuesday, March 11, 2025

Access the code and data at https://github.com/andrewheiss/snoopy-spring

Statistical methods in public policy research

Chapter for the Oxford Research Encyclopedia on Public Policy

Andrew Heiss
Georgia State University
aheiss@gsu.edu

ABSTRACT This essay provides an overview of statistical methods in public policy, focused primarily on the United States. I trace the historical development of quantitative approaches in policy research, from early ad hoc applications through the 19th and early 20th centuries, to the full institutionalization of statistical analysis in federal, state, local, and nonprofit agencies by the late 20th century. I then outline three core methodological approaches to policy-centered statistical research across social science disciplines: description, explanation, and prediction, framing each in terms of the focus of the analysis. In descriptive work, researchers explore what exists and examine any variable of interest to understand their different distributions and relationships. In explanatory work, researchers ask why does it exist and how can it be influenced. The focus of the analysis is on explanatory variables (X) to either (1) accurately estimate their relationship with an outcome variable (Y), or (2) causally attribute the effect of specific explanatory variables on outcomes. In predictive work, researchers as what will happen next and focus on the outcome variable (Y) and on generating accurate forecasts, classifications, and predictions from new data. For each approach, I examine key techniques, their applications in policy contexts, and important methodological considerations. I then consider critical perspectives on quantitative policy analysis framed around issues related to a three-part "data imperative" where governments are driven to count, gather, and learn from data. Each of these imperatives entail substantial issues related to privacy, accountability, democratic participation, and epistemic inequalities—issues at odds with public sector values of transparency and openness. I conclude by identifying some emerging trends in public sector-focused data science, inclusive ethical guidelines, open research practices, and future directions for the field.

KEYWORDS public policy; quantitative analysis; evidence-based policy; description; explanation; estimation; causal inference; prediction; econometrics

The world is awash in data. The World Bank and the Organization for Economic Cooperation and Development (OECD) offer hundreds of country-level social and economic indicators, and thousands of other measures are available from the US Census,

US state, and local government agencies. Surveys like the American Community Survey, Current Population Survey, and General Social Survey provide detailed information about states, counties, census blocks, households, and individuals. Organizations like the Jameel Poverty Action Lab (J-PAL), the Center for Effective Global Action (CEGA), the US Department of Education's What Works Clearinghouse, and the Campbell Collaboration publish data and reports from thousands of quantitative evaluations of policy interventions aimed at reducing poverty, improving health, alleviating the effects of climate change, and identifying effective education practices.

Knowing how to work with and analyze this data is a core component of public policy and administration. The Network of Schools of Public Policy, Affairs, and Administration (NASPAA) includes data analysis as one of its core competencies, and requires that students learn "to analyze, synthesize, think critically, solve problems and make evidence-informed decisions in a complex and dynamic environment" using both qualitative and quantitative data (NASPAA, 2025). This emphasis reflects a shift in how policy decisions are made and evaluated, moving from intuition and experience to a reliance on evidence-based empirical analysis.

Quantitative public policy research is inherently interdisciplinary. As Kraft & Furlong explain, "because public problems can be understood only through the insights of many disciplines, policy analysis draws from the ideas and methods of economics, political science, sociology, psychology, philosophy, and other scientific and technical fields" (2015, p. 115). Because of this mix of disciplinary traditions, statistical practices, goals, and terminology in policy research can vary wildly. A unified framework for understanding statistical approaches in policy research that transcends disciplinary boundaries is essential for advancing evidence-based policymaking.

This essay provides an overview of statistical methods in public policy, focused primarily on the United States. I begin by tracing the historical development of quantitative approaches in policy research, from early ad hoc applications to the full institutionalization of statistical analysis in federal, state, local, and nonprofit agencies. I then outline three core methodological approaches to policy-centered statistical research across social science disciplines: description, explanation, and prediction. For each approach, I examine key techniques, their applications in policy contexts, and important methodological considerations. I then consider critical perspectives on quantitative policy analysis, particularly concerns about bias, transparency, and the potential reinforcement of existing inequalities through data collection and analysis practices. I conclude by identifying some emerging trends and future directions for the field.

Brief history of statistics in public policy

Data has been used to inform government policies for centuries. Governments work to make their societies "legible" and understandable in order to better maintain control over their territories and respond to citizen demands. Technologies like cuneiform tablets documenting economic activity in ancient Sumer, Babylonian and Roman censuses, medieval English customs houses, European military records, and British cadastral maps all served as tools for governments to "see like a state" (Scott, 1998). Data

collection is even built into the United States Constitution, which mandates a decennial census for congressional apportionment (Article I, Section 2).

In the 1800s, publicly-available government data became more accessible to researchers, journalists, and policymakers, who began to use this data to lobby for legislative and social changes. For instance, in 1848, *New York Herald* editor Horace Greeley used data from the US Postal Service to show that congressional representatives were purposely overcharging travel reimbursements, leading to 1849 legislation prohibiting excess mileage charges (Klein, 2015). Later in the century, scholars and activists used data to demonstrate evidence of government-led discrimination against Black minorities throughout the United States. Ida B. Wells collected and analyzed newspaper reports and conducted extensive fieldwork to challenge the idea that Black lynching victims deserved the violence perpetrated against them due to poor morals, and instead provided evidence that lynchings were used to protect white social, economic, and political interests (Francis, 2014). Throughout his career, W. E. B. Du Bois analyzed government-collected economic data to visualize the large racial wealth gap in post-slavery America, offering dozens of visualizations, reports, and academic papers that informed policy debates in the late 19th and early 20th centuries (Du Bois et al., 2018).

Most of this public sector statistical work, however, was done on an ad hoc basis. Some proposed policies were backed by quantitative evidence, but decisions were largely made through experience and craft knowledge (Fleming & Rhodes, 2018). Near the end of the 20th century, Wilson (1887) called for the scientific study of government management and administration, arguing that systematic analysis would improve government efficiency. Though not explicitly statistical in nature, Wilson's work laid the groundwork for more quantitative approaches to governance. For instance, in the early 1900s, Charles Merriam pushed for increased statistical analysis in political science research, leading to the creation of research centers focused on quantitative political and policy studies at the University of Chicago in the 1920s, and the institutionalization of quantitative social science more broadly (Sylvan, 1991). Merriam encouraged statistical analysis at a federal level, consulting with several US presidents, and helping to found the Social Science Research Council, which remains a central hub for quantitative research today.

Following the Great Depression and its New Deal policy interventions, and especially after World War II, the US federal bureaucracy expanded rapidly in scope. In response, scholars called for more systematic study of the policy process, with Harold Lasswell and Daniel Lerner advocating for the creation of an interdisciplinary field of "policy sciences" in the 1950s (Lerner & Lasswell, 1951). Lasswell's vision was for "a more muscular and integrated version of Wilson's appeal for the scientific management of government" with research that merged Merriam-style quantitative methods with "insights from sociology, economics, business, [and] law," as well as methods from physics and biology (Allison, 2006, p. 63; Lasswell, 1951). Beginning in the 1960s with the Kennedy administration, "the impulse to clarify policy options through quantification" (Allison, 2006, p. 64) rapidly snowballed as federal agencies borrowed and adapted statistical, game theoretical, and cost-benefit analytical approaches from the Department of Defense and the RAND Corporation. Philanthropic organizations like the Ford Foundation supported this transformation, providing millions of dollars in

grants for graduate training in quantitative policy analysis throughout the 1970s (Allison, 2006).

This push for systematic quantitative analysis at a federal level culminated in the establishment of the Congressional Budget Office (CBO) in 1974. Under the direction of Alice Rivlin, the CBO developed rigorous statistical methods for budget forecasting and policy impact analysis, establishing standards for non-partisan statistical analysis that have been used to estimate the costs and score the impact of all proposed federal legislation. Similar legislative and cabinet-level executive agencies were founded or reorganized after the creation of the CBO, including the Congressional Research Service, the Government Accountability Office, and a host of offices with names including "policy," "planning," "evaluation," and "administration"—each with the goal of applying quantitative methods to systematically analyze the effects of proposed or existing public policies (Weimer & Vining, 2017, p. 35). State and local governments followed suit, establishing their own policy analysis units and mandating budget forecasts and legislative scorecards. By the 1980s, policy analysis had emerged as a distinct profession.

Before the push for the rationalization of policy research in the 1970s, most public policy research relied on qualitative methods (Breunig & Ahlquist, 2014). Today, however, policy analysis in the United States is largely quantitative. The experience of the *Journal of Policy Analysis and Management* (JPAM)—the current flagship journal for policy analysis—is illustrative. Founded in 1981, JPAM initially published qualitative policy work, including case studies, comparative and historical analysis, descriptive work, and theory building—from 1981–1984, only 27% of JPAM's research articles used any sort of quantitative methods.[1] This ratio reversed in 2000, when only a quarter of JPAM's articles were explicitly qualitative, and between 2001 and 2016, 90% of articles published in JPAM used quantitative methods, including panel regression, experiments, econometric causal inference methods, simulations, and predictive modeling. JPAM's heavy statistical emphasis continues today.

Some of this turn towards quantitative work in academic publishing is a function of editorial preferences, but much of it reflects increased demand for quantitative policy analysis by policy designers and policymakers who must adhere to legal requirements for rigorous evaluations. Beginning in the late 1980s, econometricians partnered with policymakers to develop research designs to test the causal effect of policy interventions and produce measurable evidence of policy impact. Agencies and researchers partnered to run large-scale randomized control trials (RCTs), like a Department of Labor-funded job training program (LaLonde, 1986), a Department of Housing and Urban Development housing voucher program named Moving to Opportunity (NBER, 2025), and a state-funded experiment on elementary school classroom sizes named Tennessee STAR (Mosteller, 1995). The results of these RCTs coincided with developments in non-experimental causal inference work (Card & Krueger, 1994; Imbens & Wooldridge, 2009) and helped create the "credibility revolution"—where policy evaluation is expected to have a plausible causal identification strategy to demonstrate evidence of social impact—in economics and social science more broadly (Angrist & Pischke, 2010). By the early 2000s, new institutions emerged to encourage, fund, support,

[1] Figures based on author's collected data.

and catalog the growing number of quantitative and causally-focused policy and program evaluation studies, including the Campbell Collaboration, which provides a central database and accompanying meta-analyses of a wide range of policy interventions; the Department of Education's What Works Clearinghouse, which houses evidence-based studies on education interventions; and the Jameel Poverty Action Lab (J-PAL), which funds, administers, and analyzes international development and poverty interventions around the world.

The growth of evidence-based quantitative policy analysis culminated in formal codification through the Foundations for Evidence-Based Policymaking Act of 2018, which mandated that federal agencies develop evidence-building plans and systematically evaluate their programs. This legislation represents the full institutionalization of statistical methods in policy analysis, reflecting both the maturation of quantitative methodologies and the belief that "government decisions should be based on rigorous evidence and data about what works" (Results for America, n.d.) and that systematic analysis can improve governance outcomes.

Core methodological approaches

Quantitative policy analysis and evaluation merges the statistical methods of multiple fields, including political science, psychology, and economics. Across these disciplines, quantitative researchers focus on both (1) characterizing individual social phenomena and their distributions and (2) analyzing relationships between phenomena. However, this interdisciplinary blend often creates terminological confusion, as different fields use distinct vocabulary for similar estimands, variables, tests, and procedures. While *terminology* might differ, methodologists have converged on similar categorizations of the *objectives* of quantitative research objectives. Synthesizing work by Breunig & Ahlquist (2014) and Efron (2020), I propose three fundamental categories of statistical public policy analysis:

1. **Description**, where the focus of the analysis is on exploring and understanding key variables, their distributions, and their relationships (Breunig & Ahlquist, 2014; Cleveland, 1993; Tufte, 2001; Tukey, 1977)
2. **Explanation**, where the focus of the analysis is on explanatory variables (X) to either (1) accurately estimate their relationship with an outcome variable (Y), or (2) causally attribute the effect of specific explanatory variables on outcomes (Angrist & Pischke, 2008; Breiman, 2001; Efron, 2020; Morgan & Winship, 2014; Pearl & Mackenzie, 2020; Shmueli, 2010)
3. **Prediction**, where the focus of the analysis is on the outcome variable (Y) and generating accurate forecasts, classifications, and predictions from new data (Breunig & Ahlquist, 2014; Efron, 2020)

These three categories provide a useful shorthand for describing different purposes of analysis, but they are rarely mutually exclusive. Descriptive exploratory work is necessary for both explanatory and predictive analysis, and the division between causal and non-causal inference is rarely clear-cut (Esterling et al., 2025). Researchers can shift between objectives during different phases of a single study, or may pursue multi-

ple objectives simultaneously. Table 1 summarizes these three objectives and provides some interdisciplinary disambiguation for these concepts.

Description

With legal mandates to quantitatively measure the impact of policies and programs, a substantial amount of space in public policy statistics courses and textbooks is dedicated to hypothesis testing and other inferential techniques (Berman, 2007; Bueno de Mesquita & Fowler, 2021; Nowlin & Wehde, 2024; Weber, 2024). While these approaches are important, emphasizing inference and prediction prior to understanding the available data can lead to incorrect conclusions. A key component of statistical research that should be carried out before any confirmatory or inferential data analysis is *exploratory data analysis* (EDA).

First proposed by Tukey (1965, 1977), EDA is an iterative process where researchers examine their data to discover patterns, identify anomalies, check assumptions, and develop hypotheses. Exploratory techniques help researchers understand the structure of their data before imposing theoretical models (Behrens, 1997). This approach emphasizes the importance of "getting to know" the data before conducting formal statistical tests. Even if "primary research questions are handed to you on a platter" (Wickham et al., 2023, Section 10.1) and the scope of an analysis is clear from the outset, exploring the data is still crucial for understanding its quality and discovering any unexpected patterns. Descriptive EDA includes looking at raw data values; computing univariate summary statistics like means, medians, variances, standard deviations, and ranges; computing multivariate summary statistics like correlations and crosstabs; and creating data visualizations like histograms, density plots, scatterplots, bar charts, and maps (Healy & Moody, 2014).

In general, descriptive work examines distributions, patterns, and relationships in data without necessarily making causal claims. It can look at variables by themselves (i.e. just X or just Y) or at variables in the context of other variables (i.e. the general relationship between X and Y) (Alexander, 2023, Chapter 11), and it can be done as part of either inferential and predictive analysis, or as an end in and of itself. Basic descriptive statistics are incredibly common—and valuable—in policy research (Berman, 2007, p. 96). Policymakers and managers are interested in knowing accurate estimates of all sorts of basic values, like a country's average GDP, the median unemployment rate per state, the range of PM2.5 air quality levels over the course of a year in a county, or the variance in property values within a city. For instance, Chetty, Hendren, Kline, & Saez (2014) and Chetty, Hendren, Kline, Saez, & Turner (2014) use detailed administrative data on more than 40 million individuals to describe general patterns of intergenerational mobility. Instead of arguing for a causal identification strategy or employing complex predictive methods, they largely rely on basic regression models, plots, and maps[2] to illustrate different trends in mobility and inequality throughout the United States. Many large-scale descriptive projects use public data from the US Census, state records, and other sources to provide a descriptive overview of policy trends, like the American Communities Project (2025), which classifies and maps US counties into a

[2] See https://opportunityinsights.org/ to explore interactive versions of their plots and maps.

Table 1: Summary of objectives of statistical analysis

	Description	Explanation	Prediction
General question	What exists?	Why does it exist? How can it be influenced?	What will happen next?
Focus of analysis	Focus is on any variable—understanding different variables and their distributions and relationships	Focus is on X—understanding the relationship between X and Y, often with an emphasis on causality	Focus is on Y—forecasting or estimating the value of Y based on X, often without concern for causal mechanisms
Names for variable of interest	—	• Explanatory variable • Independent variable • Predictor variable • Covariate	• Outcome variable • Dependent variable • Response variable
Goal of analysis	Summarize and explore data to identify patterns, trends, and relationships	*Estimation*: Test hypotheses or theories and make inferences about the relationship between one or more X variables and Y *Causal attribution*: A special form of estimating—make inferences about the causal relationship between a single X of interest and Y through credible causal assumptions and identification strategies	Generate accurate predictions; maximize the amount of explainable variation in Y while minimizing prediction error
Evaluation criteria	—	Confidence/credible intervals, coefficient significance, effect sizes, and theoretical consistency	Metrics like root mean square error (RMSE) and R^2; out-of-sample performance
Typical approaches	Univariate summary statistics like the mean, median, variance, and standard deviation; multivariate summary statistics like correlations and cross-tabulations	t-tests, proportion tests, multivariate regression models; for causal attribution, careful identification through experiments, quasi-experiments, and other methods with observational data	Multivariate regression models; more complex black-box approaches like machine learning and ensemble models
Examples	Summarizing unemployment rates by state or county; cross-tabulations of average unemployment rates by county and demographic group; maps of unemployment rates	Explaining variation in unemployment rates (Y) using historical and contemporary economic indicators (X); testing whether a job training program (X) causes a reduction in unemployment (Y)	Forecasting future demand for unemployment benefits (Y) based on historical and contemporary economic indicators

range of different social and economic communities, centers, and enclaves, or the Distressed Communities Index, which maps dozens of different economic indicators across ZIP codes (Economic Innovation Group, 2025).

These kinds of descriptive summary statistics can form the basis for monitoring and process evaluation work (Rossi et al., 2019) and can inform policy debates and decision-making without needing more complex explanatory or predictive approaches. For example, the Congressional Budget Office provides descriptive distributional analyses of the allocation of federal resources across various population crosstabs, like employment rates across race and income levels across family sizes (Congressional Budget Office, 2025). Similarly, public health agencies provide descriptive data on trends in disease prevalence over time and geography, which empowered policymakers and the general public during the COVID-19 pandemic (Li & Yarime, 2021).

Careful exploratory and descriptive analysis serves as both a foundation for more complex statistical methods and a valuable standalone tool in policy research. By revealing patterns and relationships that might otherwise remain hidden—and by providing researchers with a better understanding of their data—descriptive statistics are important for evidence-informed analysis.

Explanation

We can use statistics to describe variables and social phenomena, but single point estimates (e.g., the average tax revenues received by a city, the average annual unemployment rate in a state, etc.) do not generally provide enough information for making decisions or conclusions about policies (Imbens, 2021). Individual values do not indicate confidence that described characteristics reflect reality, nor do they describe how much uncertainty is inherent in those estimates (Aronow & Miller, 2019, p. 124). A core element of statistical research, therefore, is *inference*, which allows researchers both to (1) quantify the uncertainty of estimated values and (2) test hypotheses about estimates' approximations of real world phenomena. Inferential approaches move beyond description and *explain* relationships between variables, providing insights into causal mechanisms that drive policy-relevant outcomes. Explanatory analysis primarily focuses on X variables—also known as explanatory or independent variables—and their relationship with an outcome or dependent variable (Y). Explanatory analysis moves from characterizing what exists to understanding why it exists and how it might be influenced.

Estimation, inference, and hypothesis testing

Explanatory analysis entails two complementary processes: estimation and hypothesis testing. Estimation involves determining the magnitude of relationships between variables, or how much one or more X variable is associated with or influences Y. The process of estimation consists of three components: an estimand, an estimator, and an estimate. Analysts first define an *estimand*, or a target quantity of interest that is based on an underlying theory to be tested (Lundberg et al., 2021). They then apply an *estimator*—a procedure, algorithm, or technique like subtracting two averages or fitting a regression model—to calculate an *estimate* of the target quantity (Little & Lewis, 2021). For example, a researcher might be interested in the relationship of county characteristics (X) on unemployment rates (Y). Their estimand (or target quantity) might be

the difference in average unemployment rates between urban and rural counties. They would then calculate the difference in means as the estimator, resulting in an estimated difference.

Hypothesis testing evaluates whether observed estimates are statistically distinguishable from chance occurrences. Classical null hypothesis significance testing (NHST) involves comparing an observed estimate to what would be expected if there were no meaningful underlying relationship in the population. Importantly, this idea of "no relationship" does not mean a value of precisely zero; rather, it represents a distribution of values that would be considered negligible or unimportant for practical purposes, determined by the variability and sample size of the data. Hypothesis testing typically produces two key outputs: confidence intervals and p-values. A confidence interval provides a range of plausible values for the true parameter, with wider intervals indicating greater uncertainty. The p-value represents the probability of observing an estimate at least as extreme as the one calculated, if the null hypothesis is true. Analysts then must decide if there is sufficient evidence that the estimated value does not fit within the null distribution. Conventionally, 0.05 is used as an evidentiary threshold—if there is a less than 5% chance that the observed estimate could fit the null hypothesis, the estimate is considered "statistically significant" and not zero.

Standard statistical techniques such as t-tests, proportion tests, chi-squared tests, and linear regression are used for both calculating estimates and for providing details to test hypotheses about those estimates. For instance, a t-test not only estimates the difference between two group means but also tests whether that difference is statistically significant. Similarly, regression coefficients provide estimates of relationships between variables—either as slopes or shifts in intercepts—while their associated test statistics allow researchers to evaluate the statistical significance of these relationships.

Linear regression is particularly ubiquitous in explanation-focused policy analysis, as it allows researchers to explore how multiple X variables simultaneously explain a single outcome. These models are often used to examine the determinants of outcomes, like the purchase of private health insurance across a range of socioeconomic characteristics (Gutierrez, 2018), the distribution of foreign aid based on a variety of donor- and recipient-country characteristics (Bermeo, 2017), the uptake in energy efficiency tax credits across individual income levels (Jacobsen, 2019), or the adoption of nonprofit accountability practices across different organizational features (Saxton et al., 2012). Other statistical techniques like random causal forests (Athey et al., 2019; Wager & Athey, 2018) can measure the relative importance of multiple X variables, providing analysts with information about the salience of possible policy levers. For instance, Aksoy et al. (2023) explore the effect of different experimental treatments on anti-LGBT attitudes. They report regression coefficients to demonstrate the effect of individual X variables in isolation and use a random forest model to report which explanatory variables have the greatest relative influence on the outcome.

Modern statistical software makes it trivial to control for multiple explanatory variables, and it can be tempting to include as many independent variables as possible. However, this is generally a poor approach to explanatory analysis, often called "garbage can" modeling (Achen, 2005). When explaining variation in a policy outcome, researchers should ensure that the explanatory variables they include are rooted in underlying theory. Moreover, adding extra control variables can lead to unexpected

mathematical outcomes due to multicollinearity, confounding, and collider bias, and researchers must take care to not include "bad controls" in their models (Cinelli et al., 2024).

Causal attribution and causal inference

General explanatory analysis allows researchers to estimate the relationships and associations between X and Y, but on their own, these techniques cannot speak to whether relationships are mechanistic or causal. If unemployment rates are significantly higher in rural counties than in urban counties, it does not imply that forced urbanization would be a useful policy intervention to improve employment. The old adage that "correlation is not causation" holds.

However, a special form of explanatory analysis can be used to attribute changes in an outcome to a specific program or policy, allowing analysts and policymakers to discuss the causal effects of interventions. In this approach, causal attribution can be defined using a metaphor of listening and responding: "X is a cause of Y if Y listens to X and decides its value in response to what it hears" (Pearl et al., 2016, pp. 5–6). This definition aligns well with the idea of interventions as levers—policymakers can develop a program or policy to improve a social outcome, and analysts can attribute how much that policy influences the variation of that outcome, providing evidence that the intervention *causes* measurable social changes.

Calculating causal estimands requires more than statistical tests—it requires an understanding of the counterfactual, or what would have happened in the absence of a policy. The potential outcomes framework formalizes this type of counterfactual thinking (Rubin, 2005). For each unit i (an individual, a county, a state, a country, etc.), we can define two potential outcomes: Y_i^1, or the outcome if unit i receives the intervention, policy, or treatment, and Y_i^0, or the outcome if unit i does not receive the intervention. The individual causal effect for unit i is the difference between these two potential outcomes, or $Y_i^1 - Y_i^0$. However, it is not possible to simultaneously observe what would happen in a state that implemented a given policy *and* what would happen in that same state at the same point in time if it did not implement that same policy. This creates the fundamental problem of causal inference: for any unit, we can observe either Y_i^1 or Y_i^0, but never both.

Since unit-level causal effects are unobservable, researchers must estimate average treatment effects. These population-level estimands allow us to quantify the impact of policies and interventions despite our inability to observe individual counterfactuals. The estimation approach relies on comparing outcomes across different units, assuming that individual variations balance out in aggregate when proper identification strategies are employed. Two estimands are common and important in policy-related research (Greifer & Stuart, 2023). The average treatment effect (ATE) is the difference between the average outcomes for treated and untreated units ($E[Y_i^1 - Y_i^0]$), and represents the expected effect of a policy or intervention across the entire population. The average treatment on the treated effect (ATT) is the conditional average difference among only treated units ($E[Y_i^1 - Y_i^0 \mid X_i = 1]$), and represents the expected effect for those that received the intervention. More simply, the ATE represents the average effect of a policy for everyone, like the effect of a job training program on the unemployment rate for the entire state population, while the ATT represents the average

effect of a policy for those who use it, like the effect of a job training program on the unemployment rate among those who participate in it.

With observational data, though, it is not possible to simply find the difference between the average outcomes for treated and untreated units. Units self-select into policies—states pass their own laws, cities develop their own programs, and individuals sign up for interventions they feel they would benefit from. As a result, observed differences between average treated and untreated outcomes suffer from selection bias, where the choice to participate in policy and the outcome are determined by confounding factors, or common causes (Huntington-Klein, 2022; Pearl et al., 2016; Pearl & Mackenzie, 2020). To address confounding and reduce selection bias, researchers interested in causal attribution employ various identification strategies, or sets of assumptions and techniques to isolate an unbiased estimate of the effect of an intervention on an outcome. These approaches fall into two broad categories: adjustment-based identification and circumstantial identification.

Adjustment-based identification In adjustment-based identification, analysts address confounding through statistical adjustment, controlling for all variables that might influence both treatment assignment and the outcome. The identification of confounding variables is typically carried out by creating structural causal models (SCMs) and drawing directed acyclic graphs (DAGs) that represent the underlying data generating process for the treatment and the outcome (Rohrer, 2018). Following the rules of *do*-calculus, analysts can use DAGs to identify sets of covariates that need to be adjusted for to eliminate confounding (Pearl et al., 2016), as well as identify "bad controls" that should not be adjusted for (Cinelli et al., 2024)—variables that, when controlled for, can actually introduce bias rather than reduce it, such as mediators (variables that lie on the causal path between treatment and outcome) and colliders (variables that are jointly caused *by* the treatment and outcome; see Knox et al. (2020)). The actual statistical adjustment can be performed in a variety of ways, such as including all required confounders as covariates in a regression model, using confounders to calculate the propensity of treatment status and then matching observations with similar probabilities, and weighting observations by the inverse of their treatment probability (or inverse probability weighting) to create balanced groups of treated and untreated units (Heiss, 2021; Hernán & Robins, 2024).

Adjustment-based methods like propensity score matching were more common in political science and economics research in the 1990s and early 2000s (Dehejia & Wahba, 1999; Smith & Todd, 2001), and are still occasionally used (Heinrich et al., 2013), but most empirical causal work in these disciplines now relies on circumstantial identification (see King & Nielsen, 2019 for critiques of matching methods, for instance). Methods using causal graphs and inverse probability weighting remain common in epidemiological and public health research (Hernán & Robins, 2024). However, recent work has called for increased use of adjustment-based approaches in empirical political science and econometrics (Blackwell & Glynn, 2018; Huffman & Van Gameren, 2018), and newer econometrics-focused causal inference textbooks are structured around causal graphs (Cunningham, 2021; Huntington-Klein, 2021, 2022). While these adjustment-based methods rely on the difficult-to-test assumption that all confounders are observable and can be statistically adjusted, sensitivity analysis techniques allow researchers to test the robustness of estimates to unmeasured

confounding (Cinelli & Hazlett, 2020; McGowan, 2022) and enhance the plausibility of their causal identification.

Circumstantial identification Rather than attempt to account for all observable confounding, circumstantial identification approaches allow researchers to leverage research designs and special circumstances that create plausible exogenous variation in treatment assignments. These special circumstances can either be imposed by the researchers themselves through experimental manipulation, or can be found "in the wild" as natural or quasi-experiments. In randomized controlled trials (RCTs), researchers randomly assign units to be treated or untreated by a policy intervention and then calculate the difference in outcomes between the two groups. The randomization process eliminates any possible confounding, since the only process that influences a unit's access to the policy is the random assignment itself, not the unit's preferences or propensity to self-select. As a result, the unbiased, unconfounded average causal effect can be calculated as a basic difference in means ($E[Y \mid X = 1] - E[Y \mid X = 0]$). As discussed earlier, RCTs have been an incredibly powerful tool for producing policy evidence, and federal and state agencies regularly fund policy and program experiments—J-PAL alone hosts a database of more than 1,200 RCTs conducted in nearly 100 countries, with evidence related to education, public health, and poverty alleviation interventions.[3] RCTs are commonly seen as the "gold standard" in social scientific causal inference because of their plausible absence of confounding.

Researchers can also leverage exogenous manipulation to approximate the notion of random assignment. These approaches are particularly valuable for evaluating policies where randomized experiments would be impractical or unethical. Difference-in-differences (DiD) methods exploit policy changes that affect some units but not others, comparing outcome trends between treated and untreated groups before and after intervention. The key identifying assumption for this approach is the notion of "parallel trends"—in the absence of treatment, the difference between the treated and control groups would have remained constant over time (Goodman-Bacon, 2021). As long as this assumption holds, researchers can construct a plausible counterfactual prediction of the outcome for treated units in the absence of treatment. Card & Krueger (1994) used DiD to measure the effect of minimum wage increases on employment by comparing New Jersey (which raised its minimum wage) to neighboring Pennsylvania (which did not). Their work has since inspired a large literature of research across diverse policy domains including healthcare reforms, environmental regulations, educational interventions, and labor market policies (Roth et al., 2023). Recent methodological advances include two-way fixed effects (TWFE) models that expand the philosophy of DiD to allow for multiple time periods and staggered treatment adoption (Callaway & Sant'Anna, 2021), as well as techniques to address potential heterogeneous treatment effects over time (Sun & Abraham, 2021). Related to DiD, newer synthetic control methods (SCM) allow researchers to simulate the counterfactual trajectory of treated units by modeling the pre-treatment characteristics of untreated units (Abadie et al., 2010, 2015). Or, in other words, if one state implements a policy in a given year, the observed characteristics of other states are used to construct a synthetic version of the state, and

[3] See https://www.povertyactionlab.org/evaluations.

researchers can calculate the difference between the actual and synthetic outcomes to determine the causal effect.

Regression discontinuity design (RDD) methods rely on a special circumstances where treatment assignment is determined by an arbitrary threshold in a running variable that determines program eligibility, like income or test scores. Leverage for identification comes from the assumption that units right around the threshold are essentially the same except for their treatment status. For instance, if a poverty intervention program is only available to people earning less than 100% of the federal poverty line, people who are at 99% and 101% of the poverty line likely have very similar socioeconomic backgrounds and can be compared as if they were randomly assigned to the program. Under this assumption, researchers calculate the difference in average outcomes for units within a narrow bandwidth around the threshold (e.g., comparing those at 95–99.9% of the poverty line with those at 100.1–105%). RDD has been used for a variety of policy questions, including evaluating the effects of educational interventions that use test score cutoffs (Angrist & Lavy, 1999), age-based eligibility for government programs (Card & Shore-Sheppard, 2004), and geographic boundaries that determine exposure to political advertisements and their effect on voter turnout (Keele & Titiunik, 2015).

One final circumstantial approach is use an instrument—or a completely exogenous source of variation—that influences treatment assignment but influences the outcome *only through* its effect on treatment (Angrist & Pischke, 2008). Researchers use the variation in the instrumental variable (IV) to account for the endogeneity or unobserved confounding in the relationship between treatment and outcome, thus resulting in a plausible causal effect. This approach was popular in the 1990s and early 2000s, with research using proximity to college as an instrument for education to estimate the effect of education on lifetime earnings (Card, 1995), or using rainfall as an instrument for economic growth to estimate the effect of development on civil conflict (Miguel et al., 2004). However, satisfying the requirements of a valid instrument—that it is relevant (strongly correlated with the treatment), exclusive (correlated with the outcome only through the treatment), and exogenous (not correlated with any omitted variables)—has proven difficult. For example, Mellon (2024) identifies nearly 200 violations of the exclusion assumption for weather-based instruments like rainfall. Convincing instruments are increasingly hard to come by.

Circumstantial identification has seen rapid methodological innovation since the early 2000s, particularly in applied econometric research (Angrist & Pischke, 2008, 2015). The 2019 and 2021 Nobel Memorial Prizes in Economic Sciences were awarded to researchers dedicated to circumstantial identification: Abhijit Banerjee, Esther Duflo, and Michael Kremer in 2019 for their work on RCTs; and David Card, Joshua Angrist, and Guido Imbens in 2021 for their work on quasi-experimental research designs. These awards highlight the effect these methods have had on evidence-based policy research and on our understanding of causal inference in social science more broadly.

Prediction

While explanatory analysis explores the associations and effects of individual explanatory variables (X) on outcomes (Y) to determine why phenomena exist and how they can be influenced, predictive analysis focuses on what will happen next by either fore-

casting continuous outcomes and classifying categorical outcomes. Prediction is far less common than explanation in academic policy research—the majority of articles published by *JPAM* and other policy analysis journals answer explanatory questions, generally using causal inference, to explore the effects of specific policy interventions. However, forecasting and classification play crucial roles in practical governance and policy implementation.

Forecasting involves predicting numeric values. The Congressional Budget Office forecasts the economic costs of proposed federal legislation, the Federal Reserve publishes projections of inflation and GDP, states estimate future tax revenues to inform budget planning (Dadayan, 2024; McNichol, 2014). Public health agencies forecast disease spread and healthcare utilization, while transportation departments predict traffic flow and urban congestion (Hoque et al., 2021). Social service agencies use prediction to anticipate state welfare caseloads (Gurmu & Smith, 2008; Nadal-Fernandez et al., 2025) and unemployment insurance claims (Chatterji et al., 2022). Classification, meanwhile, assigns units to discrete categories based on their observable characteristics. Credit risk assessment models classify loan applicants by default risk (Meursault et al., 2024), school districts deploy early warning systems to identify students at risk of dropping out (Bird et al., 2024), law enforcement agencies use predictive policing algorithms to allocate patrol resources based on forecasted crime patterns (Lau, 2020), and criminal justice agencies set bail based on risk categorization (Berk et al., 2021). These forecasts inform resource allocation decisions and policy design and are central to daily public management.

The statistical techniques for prediction differ substantially from those used for explanation (James et al., 2021). While explanatory modeling focuses on parameter estimation and hypothesis testing, predictive modeling focuses on accuracy and performance on new, unseen data. Predictive modeling typically involves partitioning data into training and testing sets. Analysts fit a model on the training data, then evaluate its performance on the reserved test data to assess generalizability. Models range in complexity. Standard regression models can be used to generate predictions of Y by plugging test data into an estimated model, but more sophisticated models tend to yield better predictions. For instance, time series methods like ARIMA and exponential smoothing models account for temporal patterns in data (Hyndman & Athanasopoulos, 2021), while machine learning approaches—including decision trees, random forests, support vector machines, and neural networks—can capture complex, nonlinear relationships without requiring explicit specification of functional forms (Athey & Imbens, 2019).

Unlike explanatory models, where researchers interpret coefficients to understand the effect of individual X variables, the parameters in many predictive models (particularly machine learning approaches) are often uninterpretable "black boxes" (Athey & Imbens, 2019). Instead of checking if coefficients are statistically significant or robust to different model specifications, researchers assess the performance of these models based on their predictive accuracy. For example, cross-validation procedures and metrics like the root mean squared error (RMSE) and the area under the receiver operating characteristic curve (AUC-ROC) provide common measures of model performance.

Despite their practical value, predictive techniques are underrepresented in policy analysis research. However, recent methodological innovations have begun to bridge

the gap between prediction and explanation. Athey & Imbens (2019) call for increased use of machine learning in econometric and policy research and provide an overview of how these techniques can complement more explanatory work. Ongoing work in sociology and econometrics seeks to combine the predictive power of machine learning while also estimating causal effects and understanding mechanisms (Brand et al., 2023; Semenova & Chernozhukov, 2021; Wager & Athey, 2018), essentially allowing for X-focused work using methods designed for predicting Y.

The pitfalls of counting, gathering, and learning from public data

This abundance of high quality data and rigorous descriptive, explanatory, and predictive methods provides policy researchers with ample evidence and tools to test theories, evaluate policies, and refine public and nonprofit programs. However, the collection and analysis of data by governments has also faced significant criticism. Fourcade & Healy (2024) describe the emergence of a three-part "data imperative," where the public and private sectors are both driven by social pressures to *count*, *gather*, and *learn* from data. As noted earlier, governments have long sought to count the social, political, economic, and demographic activities that occur under their purview. Yet, when deciding which phenomena to count and how to count them, government data collection processes can conflict with values like democratic responsiveness, leading to biased and potentially harmful results.

In the late 1800s, as both private firms and government agencies collected more records about customers and citizens, organizations sought to systematize and order this data. Insurance companies, financial firms, real estate lenders, and government benefits agencies used observable individual-level data to organize people into aggregated categories related to health, financial risk, and social status. Specific types of readily measurable individual characteristics—such as income, sex, race, ethnicity, occupation, and education—became standardized and were used both to increase private sector profits and enhance state control by making the population "legible" (Scott, 1998). Throughout the 1900s, broad social and economic indicators like poverty measures and gross domestic product (GDP) went through a similar process of aggregation and standardization (Karabell, 2014). The imperative to count continues today, as seen throughout this essay.

This pursuit of legibility, however, has created a false "impression of precision and order" (Fourcade & Healy, 2024, p. 71), where data-based policies and decision-making can *feel* systematic, scientific, and objective while failing to account for individual heterogeneity. Ordering society into easily observable categories inherently privileges certain types of measurable characteristics by flattening more nuanced details about individuals into homogenous categories. Scott (1998) argues that this oversimplification of society into quantifiable numbers led to the erasure of local knowledge and the imposition of top-down policies disconnected from local realities, often with disastrous consequences.

One commonly proposed solution for restoring local expertise to data collection has been to democratize the process for deciding what to count and allow for public partic-

ipation in social scientific work (Kitcher, 2001). Doing so arguably allows for greater diversity in the values that get embedded in social indicators. For example, inflation in the United States is measured with the Consumer Price Index, which tracks the relative prices of items in a basket of goods that reflect typical household needs. The components of the CPI are countable, legible, and measurable, but they are also laden with values pertinent to specific segments of society—the basket of goods used in the index represents the consumption habits of wealthier households (Thoma, 2024, p. 7). Policy makers then use measures like the CPI as objective indicators of economic health and create corresponding policies that privilege those who are reflected in the indicator. Thoma (2024) argues that aggregate, seemingly-objective indicators like the CPI are anti-democratic. It is possible to democratize the process of creating these indicators—for instance, people from other socioeconomic backgrounds could suggest other unmeasured CPI components that better reflect their lived reality. But even with this kind of citizen input, epistemic inequalities remain—marginal changes to components of an aggregate measure still imply that the concept being counted by the measure is important and valuable. Individuals and social groups who might not benefit from or who disagree with this data collection are left out of the process (Thoma, 2024), and data continues to be counted without them.

Accompanying the imperative to *count* social data is the imperative to *gather* as much data as possible. This injunction to collect data has been driven by multiple factors. First, it is helpful for a state to be able to see across agencies and units (e.g., government benefits offices need income information held by tax agencies), and gathering details about individuals can improve citizen experiences with government services—linking databases across agencies was a key goal of improving digital governance in the 2010s (Noveck, 2015). Beyond the intentional work of collecting data, some of the rise in data collection has been a factor of sociological isomorphism and the reduced cost of gathering and storing data. Institutions observe how more prominent organizations collect detailed data and do the same, following a mimetic logic of collecting more data "because that's what leaders in our field do—and so they must be good" (Fourcade & Healy, 2024, p. 78). Moreover, the cost of collecting and storing exceptionally detailed data has decreased substantially over the past decades, and data collection platforms have made it easy for both organizations and individual policy researchers to pick up incidental data about people. For instance, by default, survey platforms include a surprising amount of metadata about respondents including IP addresses, geolocation data, and time spent on the survey. With additional website analytics, it is possible to identify even more identifiable data, such as hardware addresses, the URLs of websites that referred users to the survey, physical home addresses, and other details. Researchers and analysts tend to collect and retain this data "just in case," with the hope that it may someday prove to be useful. Intelligence agencies like the National Security Agency have partnered with private data brokers since 2007 to collect massive amounts of information on US residents (Savage, 2024), most of which reside unused in data warehouses. While these massive repositories of data—often collected mimetically and with no explicit purpose—can be can be stored relatively cheaply, securing them against data breaches and cyberattacks and making them safely accessible to the public poses substantial liability and costs. For instance, in 2020, a large-scale attack in the United States targeted personnel data housed by the Office of Personnel Man-

agement, the Pentagon, and the State, Treasury, and Justice Departments (Sanger et al., 2020). These threats to government-gathered data underscore the risks of unchecked data accumulation and point to a need for more deliberate and purpose-driven public sector data collection practices.

The final mandate after the imperatives to count and gather is the imperative to *learn* from the collected data. As we will explore below, there are many good, scientifically sound, and equitable methods for describing, explaining, and predicting social phenomena related to the public sector. At the same time, though, "learning from data at any scale and scope is easy to do badly" (Fourcade & Healy, 2024, p. 88). As discussed earlier, data counted and gathered by governments does not always reflect population characteristics and inherently encodes epistemic inequalities that favor social majorities (Thoma, 2024). A rich literature demonstrates that analyses based on this underlying data are also systematically biased against women, racial and ethnic minorities, and disabled communities (Broussard, 2023; Criado-Perez, 2020; D'Ignazio & Klein, 2020). These biases often stem from historical underrepresentation in data collection, the use of flawed proxies for complex social phenomena, and the epistemic prioritization of majority perspectives in defining what counts as valid data.

More concerning for policymakers is the fact that policy decisions are often made based on analyses of this biased data without recognition of that bias. Instead, researchers and practitioners often assume that since quantitative policy research is based on hard numbers, it is inherently objective and bias-free (O'Neil, 2016). This is especially common with more advanced black-box predictive modeling systems, which typically lack regulation and scrutiny over the resulting predictions. For instance, during the first few months of the COVID-19 pandemic, algorithms used by state unemployment benefits offices incorrectly categorized thousands of applicants as ineligible for unemployment insurance support because of minor inconsistencies in their data (Pahlka, 2023). Misclassification can have more serious legal consequences too. Eubanks (2019) describes how many state-run child protective services agencies have turned to machine learning-based systems for predicting child abuse, where automated predictions for possible domestic abuses in households trigger automated actions by state agencies and impose automated surveillance and behavioral requirements with strict legal consequences for noncompliance. This process has few humans in the loop to oversee possible errors, and families suffer from false positives flagged by the system.

Similar biased automated policy outcomes abound. In 2024, several US cities cancelled their contacts with the ShotSpotter gunfire detection and classification algorithm following research that demonstrated that it was both racially biased and ineffective at increasing arrests, reducing crime, or detecting gun violence (Doucette et al., 2021). Other automated systems misidentify Black criminal suspects (Angwin et al., 2016; Broussard, 2023), set higher bail amounts for Black defendants (Angwin et al., 2016; Koepke & Robinson, 2018), automatically flag trans people at security checkpoints (Costanza-Chock, 2018), prefer white male graduate school applicants (Burke, 2020), offer better loans to white male lenders (Miller, 2020), and are less likely to recommend hiring non-white, male job applicants (Jaser et al., 2022). Following Thoma (2024), these methods reflect the epistemic inequality inherent in the underlying data.

Dealing with the bias and incompleteness of these predictive models is made more difficult due to the complexity and opaqueness of their statistical methods. At their core,

predictive methods use specialized statistics to recognize patterns and make guesses about future events based on those patterns. These algorithms and models construct their own "sense" of the world—similar to how governments seek to make society "legible" (Scott, 1998)—but lack human context for why specific patterns exist in the first place. Human, street-level bureaucrats can recognize why race, gender, disability, and other personal characteristics might influence someone's interaction with the government and can make personalized accommodations as needed (Alkhatib & Bernstein, 2019), but algorithms cannot. Predictive models create a sort of "average" flattened utopia based on incomplete training data where the world is legible to algorithms. As a result, people who do not fit the model's sense of the world are flagged as anomalies, judged, and punished (Alkhatib, 2021). Compounding the issue, the creators of these systems purposely market their products with overly-ambitious outcomes—criticized by some as modern "snake oil" (Narayanan & Kapoor, 2024)—many public agencies are lured into using these products, leading to worse outcomes for the public.

Statistical methods offer useful tools for understanding social phenomena and evaluating interventions, but these same tools can reinforce existing inequalities when applied uncritically. The imperatives to count, gather, and learn from data have pose challenges for democratic accountability. Policy researchers must rely on ethical frameworks that center human values, recognize the politics of measurement, and remain attentive to voices traditionally marginalized in data collection and analysis (OECD, 2021).

Future directions

Statistical work has long been a key component of public policy research, and will continue to play an important role in governance in the future. In conjunction with the modern emphasis on program and policy evaluation, the public sector turned toward digital governance in the 2010s. In 2011, Prime Minister David Cameron established the United Kingdom's Government Digital Service (GDS) unit, charged with "setting, leading and delivering the vision for a modern digital government" (Government Digital Service, 2025). Many other countries copied this approach by creating similar units and offices, including the United States Digital Service (USDS), established by President Barack Obama in 2014. These special units' missions are designed primarily to modernize outdated government systems and improve constituent experiences with government services—e.g. updating COBOL codebases first written in the 1960s, conducting user experience research on how applicants to benefits programs move through the system, and rapidly fixing the HealthCare.gov website that accompanied the Affordable Care Act (Pahlka, 2023). As a part of their missions to digitize public sector services, these units also encouraged more modern forms of data analysis and statistical work. In 2015, President Obama appointed the first United States Chief Data Scientist in the Office of Science and Technology Policy, with the mission to "responsibly unleash the power of data to benefit all Americans" (Honey, 2016). Under this mandate, USDS and OSTP encouraged data sharing across federal, state, and local agencies and supported modern and open analysis pipelines of this data, using both standard statistical tech-

niques and more advanced machine learning approaches to help policymakers make data-driven decisions.

Since the mid-2010s, advocacy groups, activists, think tanks, and public policy schools have embraced and encouraged this governmental turn toward open data practices. Much of this work targets ethics: Tauberer (2014) outlines 14 principles for open government data, including accessibility practices, the use of open formats, commitment to public input, and the importance of citizen privacy, while the Urban Institute (Urban Institute, 2025), the IAPP (IAPP, 2025), and other associations publish reports and guidelines for best data and analytics practices and lobby for digital governance policies. Policy schools have created undergraduate and graduate courses, certificates, and degree programs in policy analytics, where the techniques and methods from the field of data science are applied to issues specific to the public and nonprofit sectors. The Data Science for Public Service Consortium[4] comprises a community of dozens of public affairs programs with policy analytics curricula and provides a space for sharing teaching materials on key statistical and quantitative methods, including a set of modern data science competencies (Data Science for Public Service Consortium, 2022) like data visualization, causal inference, predictive modeling, data mining, spatial analytics, and other skills like report automation, programming, data cleaning, and project management.

This emphasis on data science—rather than basic statistics—in public policy education parallels broader movements toward open research practices in academia and transparency in policy analytics. Academic journals increasingly require that authors share their data and code as a condition for publication, and funding agencies mandate data management plans that allow for broad accessibility and reuse. Researchers can post pre-registered hypotheses, pre-analysis plans, code, and data at the Center for Open Science's Open Science Framework (OSF), the Experiments in Governance and Politics (EGAP) initiative, and the American Economic Association's Randomized Controlled Trial Registry (The Abdul Latif Jameel Poverty Action Lab, 2025). These practices not only enhance the rigor of policy research (Andrews & Kasy, 2019; Banerjee et al., 2020; Field et al., 2020) but also democratize access to evidence, allowing a wider range of stakeholders to engage with, evaluate, reuse, replicate, and expand policy data and analysis.

Statistical work in public policy is not simply a technical exercise—it is fundamentally value-laden and political. As seen throughout this essay, as governments work to "see like a state" the imperatives to count, gather, and learn from data can conflict with democratic values. The choice of what to measure, how to classify observations, which analytical techniques to apply, and how to interpret results all reflect implicit value judgments that shape policy outcomes. Concerns about data and algorithmic bias in criminal justice, healthcare, education, and social services highlight how seemingly objective statistical methods can reproduce and amplify societal inequalities when deployed without critical reflection. Future work in policy-oriented data analysis must merge technical rigor with ethical awareness. The USDS motto to "responsibly unleash the power of data to benefit all" captures this aspiration well. Statistical work should

[4] See https://ds4ps.org/consortium/

be driven by and accountable to public service values. As statistical methods continue to evolve and public data grows ever more abundant, maintaining this public service orientation is essential to ensure that policy research contributes to more effective, equitable, and democratic governance.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510. https://doi.org/10.1111/ajps.12116

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327–339. https://doi.org/10.1080/07388940500339167

Aksoy, C. G., Carpenter, C. S., De Haas, R., Dolls, M., & Windsteiger, L. (2023). Reducing sexual orientation discrimination: Experimental evidence from basic information treatments. *Journal of Policy Analysis and Management*, *42*(1), 35–59. https://doi.org/10.1002/pam.22447

Alexander, R. (2023). *Telling stories with data: With applications in R*. Chapman and Hall/CRC.

Alkhatib, A. (2021). To live in their utopia: Why algorithmic systems create absurd outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445740

Alkhatib, A., & Bernstein, M. (2019). Street-level algorithms: A theory at the gaps between policy and decisions. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 530:1–530:13. https://doi.org/10.1145/3290605.3300760

Allison, G. (2006). Emergence of schools of public policy: Reflections by a founding dean. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford Handbook of Public Policy* (pp. 58–79). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548453.003.0003

American Communities Project. (2025). *About*. https://www.americancommunities.org/about/

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–2794. https://doi.org/10.1257/aer.20180310

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575. https://doi.org/10.1162/003355399556061

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aronow, P. M., & Miller, B. T. (2019). *Foundations of agnostic statistics*. Cambridge University Press. https://doi.org/10.1017/9781316831762

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L., Olken, B., & Sautmann, A. (2020). *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics* (w26993; p. w26993). National Bureau of Economic Research. https://doi.org/10.3386/w26993

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Berman, E. (2007). *Essential statistics for public managers and policy analysts* (2nd ed.). Sage.

Bermeo, S. B. (2017). Aid allocation and targeted development in an increasingly connected world. *International Organization*, *71*(4), 735–766. https://doi.org/10.1017/S0020818317000315

Bird, K. A., Castleman, B. L., & Song, Y. (2024). Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, pam.22569. https://doi.org/10.1002/pam.22569

Blackwell, M., & Glynn, A. N. (2018). How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review*, *112*(4), 1067–1082. https://doi.org/10.1017/S0003055418000357

Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, *49*(1), 81–110. https://doi.org/10.1146/annurev-soc-030420-015345

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3). https://doi.org/10.1214/ss/1009213726

Breunig, C., & Ahlquist, J. S. (2014). Quantitative methodologies in public policy. In I. Engeli & C. R. Allison (Eds.), *Comparative Policy Studies* (pp. 109–129). Palgrave Macmillan UK. https://doi.org/10.1057/9781137314154_6

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. The MIT Press.

Bueno de Mesquita, E., & Fowler, A. (2021). *Thinking clearly with data: A guide to quantitative reasoning and analysis*. Princeton University Press.

Burke, L. (2020, December 13). The death and life of an admissions algorithm. *Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of labour market behavior: essays in honour of John Vanderkamp*. University of Toronto Press.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*(4), 772–793.

Card, D., & Shore-Sheppard, L. D. (2004). Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansions on low-income children. *Review of Economics and Statistics*, *86*(3), 752–766. https://doi.org/10.1162/0034653041811798

Chatterji, P., Han, Y., Lahiri, K., Pang, J., & Yin, Y. (2022). *Inter-state variation in disability applications during the COVID-19 pandemic* (Center Paper NB22-02). National Bureau of Economic Research. https://www.nber.org/programs-projects/projects-and-centers/retirement-and-disability-research-center/center-papers/nb22-02

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553–1623. https://doi.org/10.1093/qje/qju022

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, *104*(5), 141–147. https://doi.org/10.1257/aer.104.5.141

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cleveland, W. S. (1993). *Visualizing data*. AT&T Bell Laboratories.

Congressional Budget Office. (2025). *Distributional Analysis*. https://www.cbo.gov/about/distributional-analysis

Costanza-Chock, S. (2018). Design justice, A.I., and escape from the matrix of domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426

Criado-Perez, C. (2020). *Invisible women: Exposing data bias in a world designed for men*. Vintage.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://mixtape.scunning.com/

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.

Dadayan, L. (2024). *Beyond the crystal ball: State revenue forecasts before, during, and after the COVID-19 pandemic*. Urban Institute. https://www.urban.org/research/publication/beyond-crystal-ball-state-revenue-forecasts-during-and-after-covid-19-pandemic

Data Science for Public Service Consortium. (2022). *Data science competencies*. https://ds4ps.org/assets/data-science-competencies.pdf

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Doucette, M. L., Green, C., Necci Dineen, J., Shapiro, D., & Raissian, K. M. (2021). Impact of ShotSpotter technology on firearm homicides and arrests among large metropolitan counties: A longitudinal analysis, 1999–2016. *Journal of Urban Health*, *98*(5), 609–621. https://doi.org/10.1007/s11524-021-00515-4

Du Bois, W. E. B., Battle-Baptiste, W., & Rusert, B. (2018). *W.E.B. Du Bois's data portraits: Visualizing Black America*. Princeton Architectural Press.

Economic Innovation Group. (2025). *Distressed Communities Index*. https://eig.org/distressed-communities/

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115*(530), 636–655. https://doi.org/10.1080/01621459.2020.1762613

Esterling, K. M., Brady, D., & Schwitzgebel, E. (2025). The necessity of construct and external validity for deductive causal inference. *Journal of Causal Inference*, *13*(1), 20240002. https://doi.org/10.1515/jci-2024-0002

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador St. Martin's Press.

Field, S. M., Wagenmakers, E.-J., Kiers, H. A. L., Hoekstra, R., Ernst, A. F., & Van Ravenzwaaij, D. (2020). The effect of preregistration on trust in empirical research findings: results of a registered report. *Royal Society Open Science*, *7*(4), 181351. https://doi.org/10.1098/rsos.181351

Fleming, J., & Rhodes, R. (2018). Can experience be evidence? Craft knowledge and evidence-based policing. *Policy & Politics*, *46*(1), 3–26. https://doi.org/10.1332/030557317X14957211514333

Fourcade, M., & Healy, K. (2024). *The ordinal society*. Harvard University Press. https://doi.org/10.4159/9780674296688

Francis, M. M. (2014). *Civil rights and the making of the modern American state*. Cambridge University Press. https://doi.org/10.1017/CBO9781139583749

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Government Digital Service. (2025). *About us*. GOV.UK. https://web.archive.org/web/20250123183423/https://www.gov.uk/government/organisations/government-digital-service/about

Greifer, N., & Stuart, E. A. (2023). *Choosing the causal estimand for propensity score analysis of observational studies* (arXiv:2106.10577). arXiv. https://doi.org/10.48550/arXiv.2106.10577

Gurmu, S., & Smith, W. J. (2008). Estimating and forecasting welfare caseloads. In J. Sun & T. D. Lynch (Eds.), *Government budget forecasting: Theory and Practice* (pp. 187–222). Routledge.

Gutierrez, C. M. (2018). The institutional determinants of health insurance: Moving away from labor market, marriage, and family attachments under the ACA. *American Sociological Review*, *83*(6), 1144–1170. https://doi.org/10.1177/0003122418811112

Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, *40*, 105–128. https://doi.org/10.1146/annurev-soc-071312-145551

Heinrich, C. J., Mueser, P. R., Troske, K. R., Jeon, K.-S., & Kahvecioglu, D. C. (2013). Do public employment and training programs work? *IZA Journal of Labor Economics*, *2*(1), 6. https://doi.org/10.1186/2193-8997-2-6

Heiss, A. (2021). Causal Inference. In F. Urdinez & A. Cruz (Eds.), *R for Political Data Science: A Practical Guide* (1st ed., pp. 235–274). Chapman and Hall/CRC. https://doi.org/10.1201/9781003010623

Hernán, M. A., & Robins, J. M. (2024). *Causal inference: what if*. Chapman and Hall / CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Honey, K. (2016, February 5). *Open data: Empowering americans to make data-driven decisions*. Whitehouse.gov. https://obamawhitehouse.archives.gov/blog/2016/02/05/open-data-empowering-americans-make-data-driven-decisions

Hoque, J. M., Erhardt, G. D., Schmitt, D., Chen, M., & Wachs, M. (2021). Estimating the uncertainty of traffic forecasts from their historical accuracy. *Transportation Research Part A: Policy and Practice*, *147*, 339–349. https://doi.org/10.1016/j.tra.2021.03.015

Huffman, C., & Van Gameren, E. (2018). Covariate balancing inverse probability weights for time-varying continuous interventions. *Journal of Causal Inference*, *6*(2), 20170002. https://doi.org/10.1515/jci-2017-0002

Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. Chapman and Hall / CRC. https://doi.org/10.1201/9781003226055

Huntington-Klein, N. (2022). Pearl before economists: The Book of Why and empirical economics. *Journal of Economic Methodology*, *29*(4), 326–334. https://doi.org/10.1080/1350178X.2022.2088085

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd ed.). Otexts.

IAPP. (2025). *AI governance center*. https://iapp.org/about/ai-governance-center/

Imbens, G. W. (2021). Statistical significance, *p*-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, *35*(3), 157–174. https://doi.org/10.1257/jep.35.3.157

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

Jacobsen, G. D. (2019). An examination of how energy efficiency incentives are distributed across income groups. *The Energy Journal*, *40*(6), 171–198. https://doi.org/10.5547/01956574.40.6.gjac

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jaser, Z., Petrakaki, D., Starr, R., & Oyarbide-Magaña, E. (2022, January 27). Where automated job interviews fall short. *Harvard Business Review*. https://hbr.org/2022/01/where-automated-job-interviews-fall-short

Karabell, Z. (2014). *The leading indicators: A short history of the numbers that rule our world*. Simon & Schuster.

Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*(1), 127–155. https://doi.org/10.1093/pan/mpu014

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Kitcher, P. (2001). *Science, truth, and democracy* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/0195145836.001.0001

Klein, S. (2015). What a 166-year-old data-driven story can teach journalists today. *I/S: A Journal of Law and Policy for the Information Society*, *11*(1), 1–12. https://kb.osu.edu/handle/1811/75411

Knox, D., Lowe, W., & Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, *114*(3), 619–637. https://doi.org/10.1017/S0003055420000039

Koepke, J. L., & Robinson, D. G. (2018). Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review*, *93*(4), 1725–1807. https://digitalcommons.law.uw.edu/wlr/vol93/iss4/4

Kraft, M. E., & Furlong, S. R. (2015). *Public policy: Politics, analysis, and alternatives* (5th ed.). CQ Press.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76*(4), 604–620.

Lasswell, H. D. (1951). The policy orientation. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and method* (pp. 3–15). Stanford University Press.

Lau, T. (2020). *Predictive policing explained*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained

Lerner, D., & Lasswell, H. D. (Eds.). (1951). *The policy sciences: Recent developments in scope and method*. Stanford University Press.

Li, V. Q. T., & Yarime, M. (2021). Increasing resilience via the use of personal data: Lessons from COVID-19 dashboards on data governance for the public good. *Data & Policy*, *3*, e29. https://doi.org/10.1017/dap.2021.27

Little, R. J., & Lewis, R. J. (2021). Estimands, Estimators, and Estimates. *JAMA : The Journal of the American Medical Association*, *326*(10), 967–968. https://doi.org/10.1001/jama.2021.2886

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

McGowan, L. D. (2022). tipr: An R package for sensitivity analyses for unmeasured confounders. *Journal of Open Source Software*, *7*(77), 4495. https://doi.org/10.21105/joss.04495

McNichol, E. C. (2014). *Improving State Revenue Forecasting: Best Practices for a More Trusted and Reliable Revenue Estimate*. Center on Budget and Policy Priorities. https://www.cbpp.org/research/improving-state-revenue-forecasting-best-practices-for-a-more-trusted-and-reliable-revenue

Mellon, J. (2024). Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable. *American Journal of Political Science*, ajps.12894. https://doi.org/10.1111/ajps.12894

Meursault, V., Moulton, D., Santucci, L., & Schor, N. (2024). One threshold doesn't fit all: Tailoring machine learning predictions of consumer default for lower-income

areas. *Journal of Policy Analysis and Management*, pam.22662. https://doi.org/10.1002/pam.22662

Miguel, E., Satyanath, S., & Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, *112*(4), 725–753. https://doi.org/10.1086/421174

Miller, J. (2020, September 18). Is an algorithm less racist than a loan officer? *The New York Times*. https://www.nytimes.com/2020/09/18/business/digital-mortgages.html

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107587991

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, *5*(2), 113. https://doi.org/10.2307/1602360

Nadal-Fernandez, J. M., Pepin, G., & Schrader, K. (2025). Strengthening work requirements? Forecasting impacts of reforming cash assistance rules. *Journal of Policy Analysis and Management*, pam.22668. https://doi.org/10.1002/pam.22668

Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

NASPAA. (2025). *Standard 5: Matching Operations - Student Learning*. Network of Schools of Public Policy, Affairs, and Administration. https://www.naspaa.org/accreditation/standards-and-guidance/standard-standard-guidance/standard-5-matching-operations

NBER. (2025). *Moving To Opportunity*. National Bureau of Economic Research. https://www.nber.org/programs-projects/projects-and-centers/moving-opportunity

Noveck, B. S. (2015). *Smart citizens, smarter state: The technologies of expertise and the future of governing*. Harvard University Press.

Nowlin, M. C., & Wehde, W. (2024). Teaching quantitative methods to students of public policy. In *Handbook of Teaching Public Policy* (pp. 168–180). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781800378117/book-part-9781800378117-22.xml

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (1st ed.). Crown.

OECD. (2021). *OECD Good Practice Principles for Data Ethics in the Public Sector* (57). OECD. https://doi.org/10.1787/caa35b76-en

Pahlka, J. (2023). *Recoding America: Why government is failing in the digital age and how we can do better*. Metropolitan Books.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J., & Mackenzie, D. (2020). *The book of why: The new science of cause and effect*. Basic Books.

Results for America. (n.d.). *Achieving the promise of the Evidence Act*. Retrieved March 6, 2025, from https://results4america.org/evidence-act-resources/

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019). *Evaluation: A systematic approach* (8th ed.). SAGE.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244. https://doi.org/10.1016/j.jeconom.2023.03.008

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sanger, D. E., Perlroth, N., & Schmitt, E. (2020, December 15). Scope of Russian hacking becomes clear: Multiple U.S. agencies were hit. *The New York Times*. https://www.nytimes.com/2020/12/14/us/politics/russia-hack-nsa-homeland-security-pentagon.html

Savage, C. (2024, January 25). N.S.A. buys Americans' internet data without warrants, letter says. *The New York Times*. https://www.nytimes.com/2024/01/25/us/politics/nsa-internet-privacy-warrant.html

Saxton, G. D., Kuo, J.-S., & Ho, Y.-C. (2012). The determinants of voluntary financial disclosure by nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, *41*(6), 1051–1071. https://doi.org/10.1177/0899764011427597

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289. https://doi.org/10.1093/ectj/utaa027

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-STS330

Smith, J. A., & Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, *91*(2), 112–118. https://doi.org/10.1257/aer.91.2.112

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

Sylvan, D. J. (1991). The qualitative-quantitative distinction in political science. *Poetics Today*, *12*(2), 267–286. https://doi.org/10.2307/1772853

Tauberer, J. (2014). *Open government data*. https://opengovdata.io/

The Abdul Latif Jameel Poverty Action Lab. (2025). *Trial registration*. https://www.povertyactionlab.org/resource/trial-registration

Thoma, J. (2024). Social science, policy and democracy. *Philosophy & Public Affairs*, *52*(1), 5–41. https://doi.org/10.1111/papa.12250

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Tukey, J. W. (1965). The technical tools of statistics. *The American Statistician*, *19*(2), 23–28. https://doi.org/10.1080/00031305.1965.10479711

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Urban Institute. (2025). *Data governance and privacy*. https://www.urban.org/expertise/data-governance-and-privacy

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Weber, J. G. (2024). *Statistics for public policy: A practical guide to being mostly right (or at least respectably wrong)*. University of Chicago Press.

Weimer, D. L., & Vining, A. R. (2017). *Policy analysis: Concepts and practice* (6th ed.). Routledge.

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (2nd ed.). O'Reilly.

Wilson, W. (1887). The study of administration. *Political Science Quarterly*, *2*(2), 197–222. http://teachingamericanhistory.org/library/document/the-study-of-administration/