

Evidence-based Policy Making

Xuchao Gao^{*}

Shuo Qi[†]

Daniel L. Millimet[‡]

June 12, 2026

IN PREPARATION FOR

Handbook of Labor, Human Resources and Population Economics

ABSTRACT

Evidence-based policy making (EBP) is often treated as a singular ideal, yet it signifies fundamentally different objectives to different stakeholders. For the researcher, EBP represents the rigorous application of the scientific method to social problems; for the policymaker, it is frequently one of many competing inputs in a high-stakes political arena. This chapter explores the inherent tension between scientific evidence and political necessity, identifying the diverse entry points where data can inform debate, as well as the pervasive junctures where political considerations—such as electoral cycles, budgetary constraints, and constituent interests—intervene. To navigate this complexity, a rigorous overview of the potential outcomes framework as the foundational language of modern causal inference is provided. It details the primary econometric methods for generating credible evidence, including randomized controlled trials, observational studies, and quasi-experimental designs. However, the path from evidence generation to implementation is fraught with challenges. The plethora of issues that arise in this transition, from threats to internal validity to the “voltage effect” seen during scaling, are examined. Finally, several comprehensive case studies that illustrate these tensions in practice, offering a realistic assessment of the potential and the pitfalls of integrating scientific evidence into the machinery of government are discussed.

^{*}Southern Methodist University. Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496, United States. xuchaog@smu.edu.

[†]William & Mary. Global Research Institute and Department of Economics, 400 Landrum Drive, William & Mary, Williamsburg, VA 23186, United States. shuoqi@wm.edu.

[‡]Southern Methodist University and IZA. Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496, United States. millimet@smu.edu.

1 Introduction

1.1 Defining Evidence-Based Policymaking

Evidence-based policymaking (EBP) has become a widely endorsed principle across government and academia, yet the term “evidence” means different things to different communities. Academics typically refer to the systematic and rigorous use of research findings to inform policy decisions, emphasizing peer-reviewed studies, formal methodologies, and replicable results. Policymakers, by contrast, often work with a broader interpretation that encompasses professional expertise, political judgment, and stakeholder input alongside academic studies (Richards 2017). For the purposes of this chapter, the focus is on evidence-based policymaking as defined by the academic tradition. This approach stands in deliberate contrast to policymaking driven primarily by ideology, political expediency, anecdotal experience, or untested assumptions. While values and political considerations inevitably shape policy choices, EBP insists that the empirical consequences of policy alternatives should be understood through systematic inquiry rather than intuition or partisan preference.

Evidence-based policymaking exists along a spectrum of analytical rigor. Correlational analyses reveal associations but cannot distinguish causal relationships due to factors such as confounding and reverse causality. Selection on observables methods, relying on matching, weighting, stratification, and regression adjustment approaches, control for observed confounders. **Quasi-experimental methods**, such as difference-in-differences, instrumental variables, and regression discontinuity designs, exploit exogenous variation to strengthen causal inference (Roth et al. 2023). **Randomized controlled trials** provide especially transparent internal validity when feasible by eliminating selection bias through random assignment, but well-designed quasi-experimental and observational studies can be equally persuasive when they exploit credible institutional variation. Recent advances in machine learning, administrative data linkages, and computational methods are further expanding the analytical toolkit, enabling more nuanced estimation of heterogeneous **treatment effects** and more accurate prediction of policy outcomes (Athey, Imbens 2019).

1.2 Relevance to Labor, Human Resources, and Population Economics

Social and economic policies directly shape human behavior and life trajectories, affecting decisions about education, employment, family formation, and retirement. Individual responses to policy incentives are heterogeneous, context-dependent, and often unfold over decades, making it difficult to predict outcomes based on theory or intuition alone. These domains are also characterized by complex interactions and equilibrium effects. A policy that helps some workers may inadvertently harm others through labor market adjustments. Minimum wage increases, for example, affect not just the wages of low-income workers but also employment levels, firm entry and exit, prices, and potentially the broader wage distribution (Autor et al. 2016; Drucker et al. 2021). Early childhood interventions influence not only the targeted children but also their families, peers, and eventually the labor markets they enter. Understanding these cascading effects requires evidence that can isolate direct impacts from indirect spillovers and general equilibrium effects.

The stakes of getting policy right are exceptionally high. Social and economic policies affect

millions of lives and represent substantial public expenditures. A training program that fails to improve employment outcomes wastes not just financial resources but also participants' time. Poorly designed pension systems can leave retirees in poverty or create unsustainable fiscal burdens. Immigration policies shape demographic trajectories and economic dynamism for generations. Moreover, there is growing public and political demand for accountability in how taxpayer resources are allocated. Governments face increasing pressure to demonstrate that social programs achieve their intended goals and represent sound investments. This accountability imperative, combined with tighter fiscal constraints in many countries, has elevated the importance of evidence in justifying, designing, and refining social and economic policies.

1.3 The Central Tension: Science versus Politics

Evidence-based policymaking rests on a fundamental tension: while it aspires to scientific rigor, policy decisions are inherently political. Evidence can illuminate trade-offs, estimate costs and benefits, and predict potential consequences, but it cannot resolve normative questions about whose interests should be prioritized. Should unemployment insurance be generous to provide adequate support, or modest to preserve work incentives? Evidence can estimate behavioral responses and fiscal costs under different scenarios, but choosing between equity and efficiency remains a value judgment that scientific methods cannot adjudicate. In all cases, there is a line where science ends and politics begins.

Moreover, even when evidence is clear about a policy's average effects, it rarely provides unambiguous guidance. Research findings are probabilistic, context-dependent, and often contested. A well-designed study might show that a job training program increases employment by five percentage points in expectation, but policymakers must still decide whether this effect justifies the cost, whether it will replicate in their jurisdiction, and how to weigh it against competing budget priorities. Different stakeholders will interpret the same evidence through different lenses, emphasizing the findings that align with their prior beliefs or interests.

Yet acknowledging these tensions does not render evidence irrelevant. While evidence cannot eliminate politics, it can elevate the quality of political debate by clarifying what is known, what is uncertain, and what the likely consequences of different choices are. Evidence can expose false claims, reveal unintended consequences, and identify policies that fail to achieve their stated goals regardless of one's values. The goal of evidence-based policymaking is not to replace political judgment with algorithmic calculation, but to ensure that when political choices are made, they are informed by the best available understanding of what those choices will actually accomplish.

1.4 Roadmap

The remainder of this chapter proceeds as follows. Section 2 discusses the theoretical foundations for evidence-based policymaking and explains why market failures, behavioral frictions, and distributional concerns create a role for empirical analysis. Section 3 describes the policy process and emphasizes the iterative interaction between evidence, institutions, and politics. Section 4 reviews the principal methodologies used to generate evidence, ranging from randomized experiments to observational and quasi-experimental designs, as well as recent advances in machine learning and policy learning. Section 5

illustrates how evidence enters actual policymaking through a series of case studies drawn from labor, education, health, and social policy. Section 6 considers the broader challenges and limitations facing evidence-based policymaking. Section 7 concludes.

2 Theoretical Foundations: Why Evidence Matters

Evidence-based policymaking rests on sound theoretical foundations. Economic theory identifies reasons why markets may fail to produce socially optimal outcomes, creating a rationale for policy intervention. However, theory alone cannot determine whether a particular market failure exists, how severe it is, or which intervention will most effectively address it. Evidence serves as the critical link between theoretical possibility and practical policy, allowing policymakers to diagnose problems accurately, design interventions that account for how people actually behave, and assess whether policies achieve their intended distributional objectives.

2.1 Correcting for Market Failures

Markets sometimes fail to allocate resources efficiently, creating a potential role for government intervention. Economic theory identifies conditions under which competitive markets break down, particularly due to externalities or information asymmetries (Stiglitz 1989), but theoretical possibility does not imply empirical necessity. Evidence determines whether a market failure actually exists in a specific context, quantifies its magnitude, and evaluates whether proposed interventions improve welfare.

Information asymmetries in health insurance markets illustrate this principle. When insurers cannot observe individuals' health risks, high-risk individuals may disproportionately purchase insurance, driving up premiums and causing low-risk individuals to exit the market, potentially leading to market collapse. Mitigating this adverse selection can generate welfare gains (Akerlof 1978; Hackmann et al. 2015). The magnitude of adverse selection, however, varies substantially across institutional contexts. Empirical evidence quantifies the extent of selection, measures the resulting welfare losses, and evaluates the effectiveness of alternative policy interventions (Einav, Finkelstein 2011; Hackmann et al. 2015).

Human capital investments illustrate a different type of market failure: externalities. Individuals may underinvest in education and training because they cannot fully capture social returns. More educated workers may raise the productivity of their peers, improve civic participation and discourse, and reduce crime, while training programs may create broader labor market benefits (Pyatt 1966; Moretti 2004; Martins, Jin 2010; Giuri, Mariani 2013). Evidence quantifies the magnitude of these externalities and identifies which populations face the largest barriers to human capital investment.

2.2 Behavioral Insights and Policy Design

People do not always make decisions that maximize their well-being, even by their own standards. They procrastinate, have limited attention and self-control, are constrained by social norms, and are influenced by how choices are framed (Thaler, Sunstein 2009). While behavioral economics provides theoretical frameworks for understanding these patterns, evidence is crucial for identifying which behavioral frictions matter most in specific contexts and for designing interventions that effectively address them.

Retirement savings policy demonstrates how behavioral evidence transforms design. Theory suggests present bias and inertia may lead to undersaving, but evidence quantifies the magnitude of these behavioral effects and tests which interventions work. Studies show that automatic enrollment dramatically increases participation rates (Madrian, Shea 2001), while default contribution rates leverage the power of inertia to improve outcomes (Thaler, Benartzi 2004). Similarly, simple interventions like sending informational letters to taxpayers increase health insurance enrollment and reduce mortality (Goldin et al. 2021) and text messages can increase savings and credit card repayments but also overdraft fees (Karlan et al. 2016; Medina 2021). Beyond nudges, evidence informs the design of policies with explicit incentives by quantifying behavioral responses. How much does extending unemployment insurance benefits reduce job search effort (Chetty 2008; Schmieder, Von Wachter 2016)? Does requiring restaurants to post caloric information on menus alter food choices (Bollinger et al. 2011; Petimar et al. 2021; Cawley et al. 2020)? Do time limits on welfare receipt increase employment or primarily increase hardship? Do disability insurance eligibility rules and benefit levels affect claiming behavior? These questions require empirical answers, behavioral elasticities in particular, to determine optimal policy design. Moreover, evidence on heterogeneous treatment effects reveals that behavioral responses vary across individuals, calling for more targeted and effective policy design.

2.3 Distributional and Equity Considerations

Policies invariably create winners and losers. Evidence illuminates these distributional consequences, which is essential for assessing whether policies align with equity objectives. Consider minimum wage policy: the central question is not only whether employment falls, but who gains and who loses. Do wage increases primarily benefit low-income families or secondary earners in higher-income households? Evidence on the incidence of benefits and costs across the income distribution reveals whether the minimum wage achieves its intended distributional objectives (Dube 2019). Similarly, higher marginal tax rates on high earners can reduce inequality but may discourage labor supply or investment (Saez 2001), generous social insurance protects against unemployment and income shocks but may reduce precautionary saving and labor force participation (Chetty 2006; Kolsrud et al. 2018), and school tracking programs that group students based on ability may help high ability children at the expense of the rest (Figlio, Page 2002; Hanushek, Wössmann 2006). Empirical estimates of key elasticities determine the magnitude of these equity-efficiency trade-offs and help identify policies that achieve distributional goals at minimal efficiency cost.

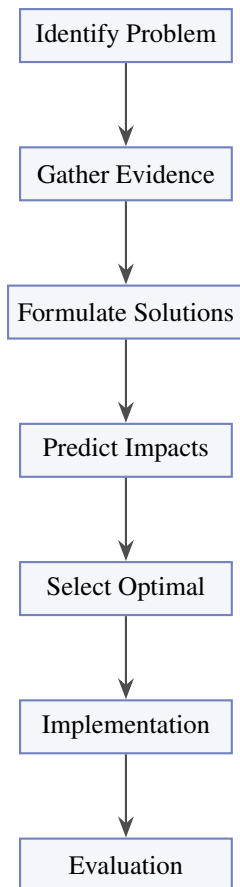
3 The EBP Process: From Question to Implementation

Evidence-based policymaking typically follows this sequence: identify a policy problem, gather relevant evidence, formulate alternative solutions, analyze their likely impacts, select the optimal policy, implement it, and evaluate the results. This cycle suggests a linear progression from problem to solution, with evidence guiding decisions at each stage as illustrated on the left side in Figure 1. However, this stylized description rarely captures how policy actually develops. In practice, the EBP process is iterative, non-linear, and deeply embedded in political and institutional contexts, where evidence competes with

interests, values, feasibility constraints, and the realities of implementation.

Problem identification begins with defining the policy issue clearly and using data to understand its scope and characteristics. Effective problem identification requires establishing the empirical magnitude of an issue, its trends over time, and its distribution across populations or regions. For instance, concerns about workforce skill gaps require data on unemployment rates by education level, employer surveys about hiring difficulties, and wage premiums for different skill sets to determine whether a genuine mismatch exists or whether other factors explain labor market outcomes (Cappelli 2015). Similarly, questions about intergenerational poverty require longitudinal data tracking economic mobility across generations to distinguish persistent disadvantage from temporary income fluctuations (Chetty et al. 2014). Data analysis helps distinguish between transitory fluctuations and persistent structural problems, identifies which populations are most affected, and provides baseline measures against which policy interventions can later be evaluated. Political priorities and media attention often determine which problems gain salience and prompt government action, but rigorous data analysis ensures that policy responses address real rather than perceived problems and that interventions are appropriately scaled.

The Stylized Model



The Iterative Reality

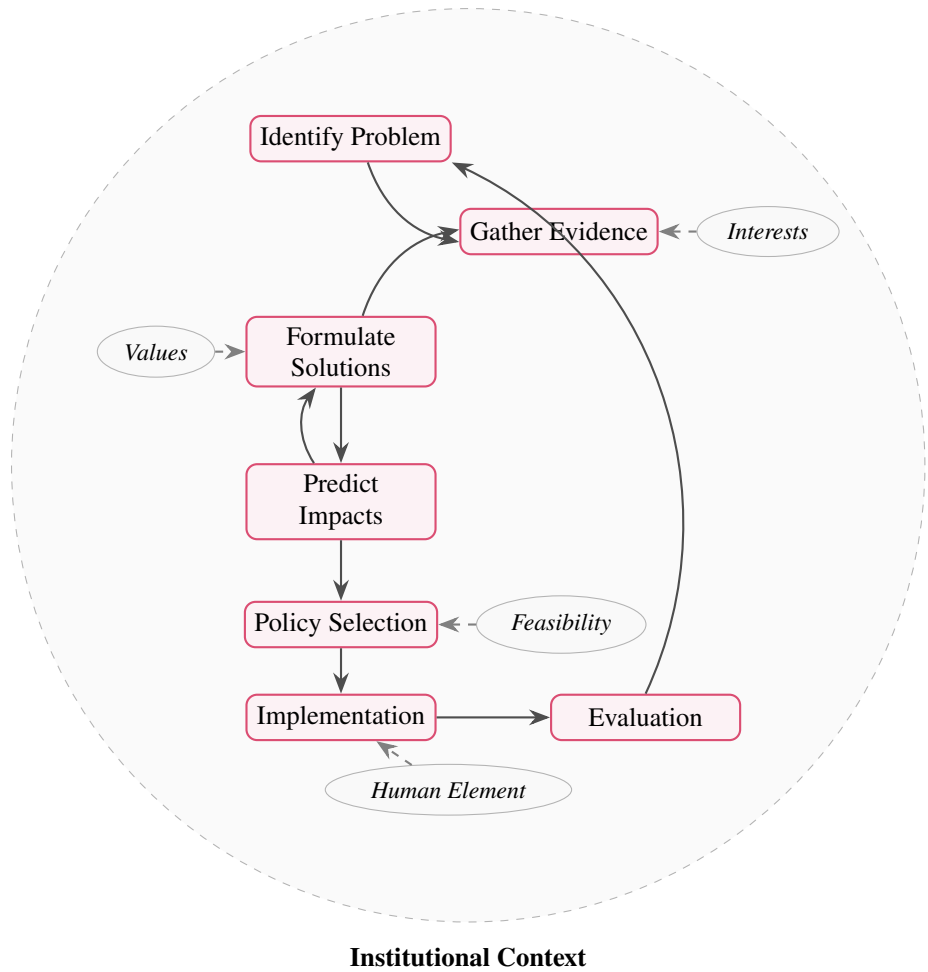


FIGURE 1. COMPARISON OF THE STYLIZED LINEAR EBP SEQUENCE AND THE ITERATIVE INSTITUTIONAL REALITY.

Formulating policy options involves drawing on existing evidence about what has worked elsewhere, economic theory about likely mechanisms, and consultations with stakeholders including practitioners, affected communities, and implementing agencies. Policymakers review prior research to identify promising approaches consistent with political values. For instance, evidence on automatic enrollment in retirement savings informed pension reforms in multiple countries (Madrian, Shea 2001), while randomized evaluations of job training programs revealed which program features improve employment outcomes (Card et al. 2018). Stakeholder engagement introduces complementary forms of knowledge: service users and affected populations possess experiential knowledge about how policies function in practice, while front-line practitioners hold professional knowledge about local contexts and implementation feasibility (Head 2010). Integrating these perspectives with research evidence requires judgment about their respective contributions. Stakeholder input primarily addresses questions of values, priorities, and feasibility, while research evidence speaks to empirical questions about effectiveness. Policymakers compare alternative policy designs by analyzing their expected costs, benefits, and distributional effects, though this process is constrained by political feasibility, budget limitations, and institutional capacity. The analysis ranges from informal reviews of existing studies to formal cost-benefit analyses and simulation models, depending on available resources and decision timelines.

Policy analysis and impact evaluation occur at multiple stages. Ex-ante analysis assesses proposed policies before implementation—answering questions about “What would happen?”—using techniques ranging from cost-benefit analysis to simulations of projected outcomes. Monitoring during implementation tracks whether programs operate as designed and whether early results align with expectations. Political differences between those involved in the policy formulation and implementation stages can create a wedge between *de jure* policy and *de facto* policy. Ex-post evaluation after implementation measures actual impacts and cost-effectiveness—answering questions about “What did happen?”. However, the iterative and non-linear nature of policymaking means these analytical stages rarely proceed sequentially. Policies evolve continuously in response to new evidence, implementation challenges, and political changes as illustrated on the right side in Figure 1. For instance, evaluations of pilot programs influence whether and how to scale interventions (Banerjee et al. 2017; Al-Ubaydli et al. 2017), and impact assessments reveal unintended consequences that prompt program redesign.

While the preceding discussion describes why evidence matters and how it enters the policy process, an equally important question concerns what constitutes credible evidence in the first place. Establishing that a policy caused an observed outcome requires more than documenting correlations; it requires a research design capable of constructing a plausible counterfactual. The methodologies discussed below differ in the assumptions they impose, the parameters they identify, and the threats to validity they confront, but all share the common objective of transforming observed data into credible causal statements.

4 Methodologies for Generating Evidence

A defining feature of EBP is its preoccupation with the counterfactual: what would have happened to a specific unit—an individual, firm, or locality—under the status quo versus an alternative policy

regime? This “fundamental problem of causal inference” is conceptual rather than merely technical (Holland 1986). Because causal effects are comparisons across mutually exclusive states of the world, and only one state is ever observed for a given unit, the realized outcome is a censored version of the underlying potential outcomes. The modern potential outcomes framework formalizes this dilemma by positing, for each unit i , a pair of potential outcomes $\{Y_i(0), Y_i(1)\}$, corresponding to untreated and treated states, respectively (Neyman 1923; Rubin 1974). The potential outcomes specify the outcome that would be realized by unit i in the state of the world without ($Y_i(0)$) and with the treatment ($Y_i(1)$). While this framework provides a universal language for empirical research, it does not, by itself, resolve the identification problem.

Despite its parsimony, the potential outcomes framework offers several deep insights. First, policy evaluation is rarely about whether an intervention “works” in the abstract, but rather how potential outcomes compare across specific states of the world. Consequently, the effects of a policy must be defined relative to a clear counterfactual alternative. Second, there is no single “policy effect”: causal effects are fundamentally unit-specific differences in potential outcomes, $Y_i(1) - Y_i(0)$. Different research designs recover distinct estimands that aggregate these unit-specific effects to answer subtly different questions. While the Average Treatment Effect (ATE) provides a population-wide summary, policymakers are often concerned with decision-relevant margins, such as the Average Treatment Effect on the Treated (ATT), the Intention-to-Treat (ITT) effect of a program offer, or Local Average Treatment Effects (LATE) that apply only to specific subpopulations shifted by an instrument or threshold (Angrist et al. 1996; Heckman, Vytlačil 2005).

Third, because unrealized potential outcomes are inherently unobserved, causal inference is never a matter of pure calculation; it requires maintained assumptions to bridge the gap between observed and counterfactual states. Fourth, the possibility of interactions between units implies that a unit’s potential outcomes may depend not only on its own treatment status but on the treatment status of others—a challenge to the Stable Unit Treatment Value Assumption (SUTVA) (Cox 1958; Rubin 1980).

In this context, an identification strategy is more than a statistical estimator; it is a structured argument linking institutional policy features to the assumptions required for causal claims. Credible designs require a well-defined source of variation, assumptions that render this variation as-if exogenous, and clarity regarding the estimand identified by the design. Because many foundational assumptions—such as SUTVA or parallel trends—are only partially testable, transparency regarding threats to validity has become a core norm of EBP rather than a peripheral technical concern.

4.1 Randomized Controlled Trials

Randomized controlled trials (RCTs) occupy a central position in the EBP canon because random assignment, in principle, severs the link between treatment status and potential outcomes. In an ideal unit-level randomized experiment, assignment is independent of both potential outcomes and potential treatment choices, ensuring comparability between treated and control groups. With full compliance in a representative sample, simple differences in mean outcomes identify average causal effects without further assumptions—a property that underlies the prominence of RCTs in policy evaluation (Duflo et al. 2007; Imbens, Wooldridge 2009).

In practice, however, ethical and institutional constraints often impede reliance on RCTs. While randomization can serve as a transparent and equitable rationing mechanism when resources are scarce, it can also raise profound ethical concerns by delaying or withholding services for vulnerable populations. These political and normative considerations fundamentally shape the feasibility of experimentation in real-world settings. Furthermore, while procedural safeguards such as pre-analysis plans and reporting standards enhance the credibility of findings, they do not resolve the inherent tradeoffs and logistical frictions of policy fieldwork (Duflo et al. 2007; Hopewell et al. 2025). Moreover, even when implementation is successful, most policy RCTs deviate from the textbook benchmark in ways that fundamentally alter both identification and interpretation.

The Challenge of Noncompliance. Noncompliance is the norm rather than the exception: eligibility or assignment is randomized, but take-up is a choice. In such settings, researchers must distinguish between two primary estimands:

- **Intention-to-Treat (ITT):** Measures the average effect of being offered the treatment. This is often the most policy-relevant estimand as it accounts for the reality that a government can offer a program but cannot compel participation.
- **Local Average Treatment Effect (LATE):** Uses the random assignment as an instrument for take-up. This recovers the causal effect specifically for “compliers”—those whose treatment status is shifted by the assignment. This requires the additional, often untestable, assumption of monotonicity (Angrist et al. 1996).

Multiple Testing and Selective Reporting. Even when randomization successfully identifies causal effects, inference can be undermined by searching across numerous outcomes or subgroups in pursuit of desired findings. Because researchers often collect a wide array of outcomes, or subjects can be divided into numerous subgroups, some estimates may appear significant purely by chance. This problem is particularly acute when researchers emphasize only favorable results, creating a distorted impression of the intervention’s effectiveness. Pre-analysis plans, which specify primary outcomes and empirical strategies before examining the data, help limit researcher discretion and reduce concerns about selective reporting. Statistical adjustments for multiple testing, such as family-wise error rate controls or false discovery rate procedures, provide additional safeguards by accounting for the increased probability of spurious findings when many hypotheses are examined simultaneously (Benjamini, Hochberg 1995; Anderson 2008; Olken 2015). These practices reinforce the broader principle that credible policy evaluation depends not only on random assignment, but also on transparent and disciplined empirical analysis.

Clustering and Inference. Policy interventions are frequently implemented at the cluster level—schools, firms, or local agencies. This design choice necessitates inference procedures that account for within-cluster correlation of errors. It also serves as a reminder that the effective sample size and statistical power are driven more by the number of clusters than the number of individuals within them. Stratifica-

tion and blocking are commonly employed to improve precision, though users must ensure the correct interpretation of the parameter estimated.

SUTVA, Spillovers, and Attrition. Departures from the ideal RCT benchmark can reintroduce the very biases randomization seeks to eliminate. Differential attrition—where the treated and control groups drop out of the study at different rates or for different reasons—can destroy the balance achieved at baseline. Furthermore, interference and **spillovers** challenge SUTVA. Especially in labor markets or public health, a policy might affect non-participants through **general equilibrium** effects or social contagion. In these cases, the potential outcome of the control group is compromised, requiring more complex designs that vary treatment intensity across markets (Miguel, Kremer 2004).

Scalability and External Validity. A recurring theme in the literature is that strong **internal validity** does not guarantee **external validity**. When interventions are scaled or transplanted, implementation fidelity often drops, and participant composition may shift. A growing body of work cautions against interpreting RCT estimates as context-free truths, emphasizing instead their conditional nature (Deaton, Cartwright 2018; Vivaldi 2020). Impacts often attenuate—or even reverse—when delivered by government bureaucracies under routine conditions rather than by researchers or specialized NGOs (Bold et al. 2013).

4.2 Observational Studies Under Selection on Observables

Observational studies under selection on observables posit that causal effects can be recovered by conditioning on a sufficiently rich set of pre-treatment covariates. The basic idea is simple: if treated and untreated units differ in ways that affect outcomes, the researcher should adjust for these differences so that treatment assignment becomes conditionally comparable, or “as-if” randomly assigned, within covariate cells (Imbens, Rubin 2015; Pearl 2009). This class of methods includes regression adjustment, matching, subclassification, weighting, propensity-score methods, and modern machine-learning-based approaches to covariate adjustment.

4.2.1 Identification Under Conditional Independence

Let $D_i \in \{0, 1\}$ denote treatment status, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, and let X_i denote a vector of observed pre-treatment covariates. The observed outcome is

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

The selection-on-observables paradigm rests on two core assumptions. The first is the Conditional Independence Assumption (CIA), also called unconfoundedness, conditional exchangeability, conditional ignorability, or backdoor admissibility. For the average treatment effect, it requires

$$(Y_i(1), Y_i(0)) \perp D_i \mid X_i.$$

This assumption states that, after conditioning on X_i , treatment status is independent of the potential outcomes. Thus, any systematic difference in observed outcomes between treated and control units with the same covariates can be interpreted as causal rather than as selection bias.

The second assumption is overlap, also called positivity or common support. Let

$$e(X_i) = \Pr(D_i = 1 \mid X_i)$$

denote the **propensity score**, the probability that a unit with characteristics X_i receives the treatment. The overlap condition requires

$$0 < e(X_i) < 1 \quad \text{almost surely.}$$

Without overlap, some subpopulations with a particular set of characteristics are deterministically assigned to treatment or control, making their counterfactual outcomes unidentified without strong parametric extrapolation.

Under CIA and overlap, the average treatment effect is identified by covariate adjustment:

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mathbb{E}(Y_i \mid D_i = 1, X_i) - \mathbb{E}(Y_i \mid D_i = 0, X_i)].$$

Similarly, the average treatment effect on the treated is

$$\tau_{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] = \mathbb{E}[Y_i - \mathbb{E}(Y_i \mid D_i = 0, X_i) \mid D_i = 1].$$

In words, the ATE or ATT is a weighted average of Conditional Average Treatment Effects (CATEs), defined as the expected difference in outcomes with and without treatment for units with characteristics X_i . The weights reflect the distribution of X_i in the population for the ATE or treatment group for the ATT. Intuitively, among units with identical X_i , treatment assignment mimics a RCT.

When X_i is high-dimensional, matching or subclassifying directly on X_i becomes practically infeasible because of the curse of dimensionality: observing units in both the treatment and control group with identical X_i becomes unlikely. Rosenbaum, Rubin (1983) resolved this problem by introducing the scalar propensity score, $e(X_i)$. The authors showed that if unconfoundedness holds conditional on X_i , then it also holds conditional on the propensity score:

$$(Y_i(1), Y_i(0)) \perp D_i \mid e(X_i).$$

This dimension-reduction result allows researchers to estimate treatment effects by matching, subclassifying, or weighting on $e(X_i)$ rather than on the full vector X_i . In practice, $e(X_i)$ is unknown and must be estimated. This first-step estimation problem affects the asymptotic distribution of matching and weighting estimators and should be accounted for in standard errors (Heckman et al. 1997; Heckman et al. 1998b).

In settings where X_i is high-dimensional, selection-on-observables approaches have integrated machine learning to relax restrictive functional-form assumptions. **Double/Debiased Machine Learning**

(DML) uses flexible prediction methods to estimate “nuisance” functions such as the outcome regression and the propensity score (Chernozhukov et al. 2018a). By doing so, DML modernizes selection-on-observables approaches by weakening functional-form restrictions and allowing richer covariate adjustment. However, it does not remove the central identifying assumption. Like matching, weighting, and regression adjustment, DML remains critically dependent on CIA and overlap.

The selection-on-observables paradigm has been the subject of intensive empirical and methodological debate. In the classic evaluation of the National Supported Work Demonstration, nonexperimental estimates were compared with an experimental benchmark, and standard nonexperimental methods were found to often fail to replicate the experimental result (LaLonde 1986). This evidence was later revisited using propensity-score matching, with the argument that the experimental benchmark could be recovered when detailed pre-treatment earnings histories were included in X_i (Dehejia, Wahba 1999; Dehejia, Wahba 2002). Subsequent work showed, however, that these results were highly sensitive to sample restrictions, covariate choices, and the construction of the comparison group (Smith, Todd 2005).

This debate highlighted the critical interaction between data quality and statistical method. Conventional evaluation bias can be decomposed into components due to support mismatch, density mismatch, and unobserved confounders (Heckman et al. 1997; Heckman et al. 1998b; Heckman et al. 1998a). A central lesson from this literature is that support mismatch and data-quality problems, such as comparing individuals from different local labor markets or from surveys with different questionnaires, can be empirically larger sources of bias than unobserved confounding itself. The literature also introduced difference-in-differences matching estimators to eliminate time-invariant sources of bias, reinforcing the broader message that no statistical estimator can compensate for poor research design.

4.2.2 Design, Balance, and Sensitivity Analysis

A foundational tenet of contemporary causal inference is that the scientific validity of an observational study is determined primarily at the design stage, before examining outcome data. The historical roots of this view go back at least to the planning of observational studies on human populations, the propensity-score design perspective, and the view that “design trumps analysis” (Cochran 1965; Rosenbaum, Rubin 1983; Rubin 2001; Rubin 2008). The core objective of an observational design is to reconstruct, as closely as possible, a hypothetical randomized experiment. This reconstruction should be conducted “blindly”, without using outcome variables, so that researchers do not search over samples, covariates, or specifications to obtain statistically significant results.

One increasingly influential way to operationalize this principle is the target trial emulation framework. Instead of beginning with an available observational dataset and asking what association can be estimated, the investigator first writes down the protocol of the hypothetical randomized trial that would answer the causal question of interest (Hern’an, Robins 2016). This protocol specifies eligibility criteria, treatment strategies, the time of treatment assignment or “zero time”, follow-up rules, outcomes, causal contrasts, and target estimands such as intention-to-treat or average treatment effects for relevant subpopulations. The observational cohort is then used to emulate this target trial as closely as possible. By aligning eligibility, treatment initiation, follow-up, and estimands with a trial protocol, target-trial emulation helps avoid structural biases, including immortal time bias and other forms of self-inflicted

bias caused by incoherent timing definitions (Hern'an et al. 2016).

The design stage focuses on two closely related tasks: achieving covariate balance and ensuring overlap, or common support. As the propensity score summarizes the observed covariate information relevant for treatment assignment under selection on observables, propensity-score distributions should be visually inspected using density plots, histograms, or empirical distribution functions. Areas in which treated units have no comparable controls, or controls have no comparable treated units, indicate a failure of overlap. Units outside the overlapping region should typically be trimmed, reweighted, or assigned to a more limited target population (Crump et al. 2009). This is not merely a technical adjustment. Trimming non-overlapping tails trades some external validity for greater internal validity by avoiding dangerous model extrapolation and reducing estimator variance.

After constructing comparable treated and control groups, balance should be assessed using standardized mean differences rather than conventional t -tests. For a scalar covariate X_k , a common diagnostic is

$$\text{SMD}_k = \frac{\bar{X}_{1k} - \bar{X}_{0k}}{\sqrt{(s_{1k}^2 + s_{0k}^2) / 2}},$$

where $\bar{X}_{1k}$ and $\bar{X}_{0k}$ are the treated and control means and s_{1k}^2 and s_{0k}^2 are the corresponding sample variances. Unlike t -tests, standardized mean differences are not mechanically driven by sample size and therefore provide a more design-oriented measure of imbalance (Austin 2009; Stuart 2010). A credible observational design should report balance before and after matching, weighting, or trimming, and should make clear which population the final design represents.

Even after careful design, the key identifying condition in many observational studies—conditional independence—remains fundamentally non-testable. Therefore, a credible design should also include a sensitivity analysis that asks how strong unmeasured confounding would need to be to overturn the study's conclusions. If only a small amount of hidden bias would eliminate the result, the design is fragile. If an implausibly large unobserved confounder would be required, the inference is more robust. Two prominent approaches are especially useful for formalizing this robustness. Because these are less well-known, detailed coverage is provided.

Rosenbaum-style sensitivity analysis and marginal sensitivity models. The classic sensitivity-analysis framework bounds the extent to which hidden confounding can affect treatment assignment among units that are comparable on observed covariates (Rosenbaum 2002). Let

$$\pi_i = \Pr(D_i = 1 \mid X_i, U_i),$$

where U_i is an unobserved confounder. For two units i and j with the same observed covariates, $X_i = X_j$, the Rosenbaum sensitivity model assumes that their odds of treatment assignment may differ by at most a factor $\Gamma \geq 1$:

$$\frac{1}{\Gamma} \leq \frac{\pi_i / (1 - \pi_i)}{\pi_j / (1 - \pi_j)} \leq \Gamma.$$

When $\Gamma = 1$, treatment assignment is as good as random within observed covariate strata. Larger values of Γ allow progressively stronger hidden bias. Researchers can report the largest value $\hat{\Gamma}^*$ for which the

qualitative conclusion remains unchanged:

$$\hat{\Gamma}^* = \sup \left\{ \Gamma \geq 1 : \begin{array}{l} \text{the null conclusion is still rejected} \\ \text{under the sensitivity analysis} \end{array} \right\}.$$

The asymptotic counterpart of this robustness threshold is the design sensitivity, which compares how different designs and test statistics perform under hidden bias (Rosenbaum 2004).

For weighting-based estimators, this idea has been extended through a marginal sensitivity model (Zhao et al. 2019). Let $e(X_i) = \Pr(D_i = 1 \mid X_i)$ denote the observed propensity score and let

$$e_0(X_i, U_i) = \Pr(D_i = 1 \mid X_i, U_i)$$

denote the full-data propensity score that would be available if the unobserved confounder were observed. The marginal sensitivity model imposes

$$\frac{1}{\Gamma} \leq \frac{e_0(X_i, U_i)/\{1 - e_0(X_i, U_i)\}}{e(X_i)/\{1 - e(X_i)\}} \leq \Gamma.$$

This model asks how much the true odds of treatment assignment could deviate from the odds implied by observed covariates alone. It is therefore especially natural for inverse-probability weighting, covariate balancing, and related designs.

Partial- R^2 sensitivity analysis. A second approach evaluates omitted-variable bias using partial R^2 measures (Cinelli, Hazlett 2020). Consider a regression adjustment of the form

$$Y_i = \tau D_i + X_i' \beta + \varepsilon_i.$$

Instead of specifying a full model for the treatment assignment mechanism or the distribution of unobserved confounders, the partial- R^2 framework asks how strongly an omitted confounder Z_i would have to be associated with both treatment and outcome, conditional on observed covariates, to explain away the estimated effect. The two key sensitivity parameters are

$$R_{D \sim Z|X}^2 \quad \text{and} \quad R_{Y \sim Z|D,X}^2,$$

where $R_{D \sim Z|X}^2$ measures the residual association between the omitted confounder and the treatment, and $R_{Y \sim Z|D,X}^2$ measures the residual association between the omitted confounder and the outcome.

The robustness value, denoted RV, is the minimum common strength of association that an omitted confounder would need to have with both treatment and outcome to reduce the estimated treatment effect to zero. A related diagnostic is the partial R^2 of the treatment with the outcome, conditional on observed covariates:

$$R_{Y \sim D|X}^2 = \frac{t_D^2}{t_D^2 + \text{df}},$$

where t_D is the t -statistic for the treatment coefficient and df is the residual degrees of freedom. This quantity can be interpreted as the treatment

effect in the extreme benchmark case where the omitted confounder explains all residual variation in the outcome. By reporting RV , $R^2_{Y \sim D|X}$, and calibrated benchmarks based on observed covariates, researchers can move debates about omitted confounding from vague speculation to quantitative assessment.

4.3 Design-Based Quasi-Experimental Methods

When randomization is infeasible or politically constrained, quasi-experimental methods exploit institutional rules, policy discontinuities, timing variation, or “natural” shocks to estimate the relevant counterfactual. Unlike RCTs, where the assignment mechanism is known by design, the credibility of quasi-experiments hinges on the institutional plausibility of their identifying assumptions and the researcher’s ability to defend the as-if randomness of the observed variation. Unlike selection-on-observables approaches, quasi-experimental methods allow at least some unobserved differences between units to be correlated with potential outcomes and treatment assignment, thereby relaxing CIA. Because these designs are discussed in detail elsewhere in this *Handbook*, the discussion here is brief.

Difference-in-Differences and Synthetic Controls. **Difference-in-Differences** (DiD; see Callaway 2020) remains the workhorse of policy analysis, identifying treatment effects by comparing temporal changes in outcomes across treated and control groups. Its validity rests on the parallel trends assumption: the requirement that, absent the intervention, the treated units would have followed the same trajectory as the control group. Importantly, DiD permits time-invariant, unobserved differences between treated and control units, provided those differences would not have generated differential trends in the absence of treatment.

Recent econometric advances have fundamentally reshaped this field. In settings with staggered treatment adoption, where different units enter the policy regime at different times, standard two-way fixed effects (TWFE) estimators can yield biased or uninterpretable weighted averages, particularly when treatment effects vary over time. This has spurred a transition toward “robust” DiD estimators that prioritize transparent estimand definition and avoid the “negative weighting” pitfalls of conventional linear models (Callaway, Sant’Anna 2021; Sun, Abraham 2021; Goodman-Bacon 2021). Less well appreciated, but equally important, is the role of functional form in the plausibility of the parallel trends assumption (Roth, Sant’Anna 2023).

Synthetic Control Methods (SCM) complement this framework, particularly when analyzing policies affecting a small number of treated units, such as a single state or country. SCM constructs a data-driven “counterfactual unit” from a weighted combination of donors, making identifying assumptions more transparent by visually demonstrating pre-intervention fit. While the method allows for some unobserved differences across treated and control units that also affect outcomes, it remains dependent on the stability of the relationship between the treated unit and its synthetic donor over the evaluation horizon (Abadie 2021).

Instrumental Variables and Regression Discontinuity. **Instrumental variables** (IV) exploit variation that affects treatment uptake but influences outcomes only through that treatment. The central identifying restriction is the exclusion restriction, together with relevance and, in many applications, monotonicity.

IV can bypass unobserved confounding in the relationship between treatment and outcomes, but it is often limited by weak-instrument concerns, difficulties in defending the exclusion restriction, and the local nature of the identified effect, commonly interpreted as a local average treatment effect.

Regression Discontinuity (RD; see Melly, Lalive 2020) designs leverage sharp or fuzzy thresholds in assignment rules based on a running variable. Their identifying logic is intuitive: units just below a threshold serve as a near-perfect counterfactual for units just above it, provided that potential outcomes vary smoothly through the cutoff and units cannot precisely manipulate their position relative to the threshold. While highly credible, RD estimates are inherently local to the threshold, often limiting their immediate external validity to the specific margin of eligibility (Lee, Lemieux 2010).

4.4 Descriptive and Contextual Evidence

While RCTs, selection-on-observables designs, and quasi-experimental methods are the primary tools for causal identification, descriptive and contextual evidence remain indispensable to evidence-based policymaking. Cross-sectional and panel data analyses are essential for documenting baseline disparities, forecasting program take-up, describing institutional settings, and elucidating the mechanisms through which policies may operate. However, as noted in the potential outcomes framework, causal interpretations of these data require strong “unconfoundedness” assumptions—positing that selection into treatment is independent of potential outcomes once observed covariates are controlled for—which are rarely fully defensible in complex policy environments (Imbens, Wooldridge 2009).

Cross-sectional comparisons frequently conflate causal effects with selection bias, a problem exacerbated by the difficulty of measuring latent variables such as motivation, social networks, administrative capacity, or political constraints. Panel data can mitigate time-invariant unobserved heterogeneity through fixed-effects estimators, but time-varying confounders and dynamic selection remain persistent threats to validity. Consequently, nonexperimental evidence is most robust when used to complement, rather than substitute for, designs with clearer sources of causal identification.

A central tenet of evidence-based policymaking is maintaining a rigorous distinction between descriptive and causal claims. Observational data are often the only means of determining a policy’s reach, its distributional impacts across subgroups, the characteristics of affected populations, and the external relevance of experimental or quasi-experimental estimates. Evidence-based policymaking is strongest when these distinct roles—causal identification via design and contextualization via descriptive evidence—are made explicit rather than blurred (Holland 1986).

4.5 Machine Learning for Prediction and Causal Heterogeneity

Machine learning (ML) has transformed EBP through two broad channels: prediction and causal inference.

- **Prediction and Targeting:** ML can improve risk prediction, program targeting, fraud detection, early-warning systems, and administrative triage. These applications are useful when policy agencies need to allocate attention or resources under capacity constraints. However, prediction is

not the same as causal evaluation; a variable that predicts need or risk does not necessarily identify the effect of an intervention.

- **Causal Forests**: These allow for the data-driven discovery of CATEs. This enables policymakers to identify which subpopulations benefit most from a policy without the risks of “p-hacking” inherent in manual subgroup analysis (Wager, Athey 2018).
- Double/Debiased ML (DML): Building on early work by Robinson (1988) and leveraging the Frisch-Waugh-Lovell theorem, DML uses ML to control for high-dimensional nuisance functions while maintaining valid frequentist inference for the treatment effect. This allows researchers to account for a vast array of potential confounders in a computationally feasible way while preserving the credibility of the underlying identification strategy (Chernozhukov et al. 2018b).

Despite their computational power, these advances do not substitute for a credible research design; rather, they rely on it. Analysts must remain cognizant that while ML can mitigate functional-form misspecification, handle high-dimensional settings, and improve the precision of estimated heterogeneity, it does not relax the fundamental requirement for an underlying identification strategy rooted in institutional reality.

4.6 From Treatment Effects to Policy Choice

A treatment effect is not yet a policy recommendation. Impact evaluation estimates how outcomes change under a policy relative to a counterfactual, but policymaking also requires a criterion for deciding whether the change is socially desirable. Moving from causal effects to policy choice requires combining impact estimates with fiscal costs, administrative costs, behavioral responses, externalities, and distributional weights. This is especially important when treatment effects are heterogeneous, when participation is voluntary, or when policy changes affect the composition of treated individuals.

- Statistical Treatment Rules and Empirical Welfare Maximization: This literature asks which individuals should receive treatment when policymakers face budget, capacity, or eligibility constraints (Manski 2004; Kitagawa, Tetenov 2018). The policy problem is to choose a treatment rule that maximizes expected social welfare subject to constraints. In this framework, treatment-effect heterogeneity is not a secondary complication; it is the basis for targeting.
- **Policy Learning**: Machine learning has expanded the practical scope of policy learning by allowing researchers to estimate heterogeneous treatment effects and construct data-driven assignment rules. Methods such as causal forests and related policy-learning approaches can use high-dimensional covariates to identify groups with larger expected gains (Wager, Athey 2018; Athey, Wager 2021). Related methods and applications are discussed in the companion handbook chapter on causal machine learning (Lechner 2023). However, optimal targeting is not equivalent to assigning treatment to individuals with the largest predicted treatment effects. Policymakers must also account for program costs, take-up behavior, administrative feasibility, equity objectives, and legal or political restrictions. Take-up is especially important for social programs because eligibility does not mechanically translate into participation (Ko, Moffitt 2024).

- **Marginal Treatment Effects:** Many policies do not assign treatment directly; they change eligibility rules, subsidies, information, or benefit generosity, thereby shifting who participates. In such cases, the relevant policy parameter is often the effect for marginal participants induced into treatment by the policy change, rather than the average effect for the whole population or for those already treated (Heckman, Vytlačil 2005; Carneiro et al. 2010). This distinction is especially important for labor, education, health, and social insurance programs, where expansions often change the composition of participants.
- **Welfare Metrics:** Recent work emphasizes that policy evaluation should account for beneficiaries' willingness to pay and the net cost to the government. The marginal value of public funds provides one way to compare policies across domains by relating beneficiaries' willingness to pay for a policy to its net fiscal cost (Hendren, Sprung-Keyser 2020). This approach is useful because many social programs generate both direct benefits and fiscal externalities, such as future tax revenue, reduced medical spending, or changes in labor supply.
- **Cost-Benefit Analysis (CBA):** CBA converts disparate impacts into a common monetary metric. These impacts must account for not only partial but also general equilibrium effects. In human capital research, this often requires extrapolating short-term impacts—such as test score gains—into lifecycle earnings, accounting for overall changes in the labor market from policies implemented at scale, a process highly sensitive to discount rates and modeling assumptions (Heckman et al. 2010). Ignoring externalities can vastly misstate the cost-benefit ratio of a policy (Miguel, Kremer 2004).
- **Equity and Distribution:** Modern CBA increasingly moves beyond aggregate efficiency to report incidence by demographic group. For instance, recent guidance emphasizes documenting how policies affect outcomes across racial and socioeconomic lines, acknowledging that the “optimal” policy depends heavily on the social welfare function chosen by the policymaker (Office of Management and Budget 2023).
- **Systematic Reviews:** Because single studies are often noisy and context-specific, meta-analyses (following PRISMA protocols) are essential for identifying stable policy patterns. Large-scale reviews in labor economics have highlighted, for example, the consistent long-term efficacy of human capital investment relative to more transient job-search assistance (Card et al. 2018).

Policy learning and welfare analysis also introduce new risks. Estimated treatment rules may overfit the data, perform poorly out of sample, or rely on variables that are unavailable, manipulable, or politically unacceptable for targeting. Welfare rankings may be sensitive to discount rates, distributional weights, assumptions about externalities, or the shadow cost of public funds. For this reason, policy recommendations based on estimated rules should be evaluated not only by predicted welfare gains, but also by robustness, transparency, fairness, and implementability.

5 Evidence in Practice

This section uses case studies from labor, education, health, and social policy to examine how empirical evidence enters actual policymaking. The goal is not to rank methods mechanically, but to show how different research designs answer different policy questions under different political and institutional constraints. Randomized controlled trials often have strong rhetorical appeal because their identifying logic is easy to communicate: treatment was assigned by chance. Yet many policy decisions cannot be randomized, either because the policy is already operating at scale, because withholding treatment would be politically or ethically infeasible, or because the relevant outcomes unfold over long horizons.

For this reason, selection on observables approaches and quasi-experimental studies have been central to evidence-based policymaking. Properly designed nonexperimental studies can be highly credible when they exploit rich administrative data, institutional rules, longitudinal histories, or repeated policy variation. They may also be more directly policy-relevant than small-scale experiments because they evaluate policies as actually implemented, including bureaucratic frictions, political targeting, heterogeneous local capacity, and general-equilibrium responses. In many domains, observational evidence affects policy not through a single decisive study, but through accumulation, synthesis, and translation into policy-facing documents such as official reports, expert reviews, regulatory assessments, and legislative analyses.

The cases below therefore compare experimental and observational evidence not as a simple hierarchy, but as alternative ways of linking science to policy decisions. The section first discusses several examples in which observational or quasi-experimental evidence shaped policy through cumulative synthesis and institutional uptake, including tobacco control, traffic safety, and nutrition assistance. It then turns to randomized and lottery-based evaluations in job training, education, teacher incentives, early childhood programs, administrative burden, and youth mentoring. Together, these cases illustrate how rigorous evidence generation can inform program evaluation, clarify mechanisms, reveal implementation constraints, and, in some instances, reshape legislation, administrative practice, and the allocation of public resources. The case studies are briefly summarized in Table 1.

5.1 Observational Evidence

Tobacco Control: Smoking and Lung Cancer. Tobacco control provides an important case in which policy-relevant evidence was built from epidemiological, clinical, and biological research. The 1964 Surgeon General’s report synthesized a large body of evidence on cigarette smoking and health, including cohort studies, case-control studies, mortality statistics, and biomedical research. The report concluded that cigarette smoking was causally related to lung cancer and other serious health risks (U.S. Department of Health, Education, and Welfare 1964). Its strength lay in the consistency of findings across different research designs, populations, and mechanisms.

The report helped establish smoking as a public health issue rather than solely a matter of individual consumption. Subsequent policy responses included warning labels on cigarette packages and restrictions on cigarette advertising, including the later ban on broadcast advertising (Centers for Disease Control and Prevention 2024). The case is therefore notable for the role of an authoritative evidence review in

translating a broad scientific literature into a policy-relevant assessment.

The main advantage of this case is the depth and convergence of the evidence base. The association between smoking and disease was documented repeatedly and supported by plausible biological mechanisms. Its limitation is that the evidence did not arise from a single clean policy experiment. As with many public health questions, causal interpretation depended on the accumulation of evidence and the assessment of alternative explanations, including selection, reporting, and other health-related behaviors.

Minimum Legal Drinking Age. Traffic safety provides a second case in which policy variation generated observational evidence relevant to later policy decisions. In the 1970s and early 1980s, states differed in their minimum legal drinking ages, partly because some states lowered the drinking age while others did not. This variation allowed researchers and agencies to examine the relationship between drinking-age laws and alcohol-related motor vehicle crashes among young drivers.

The evidence suggested that lower minimum drinking ages were associated with increases in alcohol-related crashes and fatalities among young people. These findings helped support the federal move toward a uniform minimum legal drinking age of 21 through the National Minimum Drinking Age Act of 1984 (National Highway Traffic Safety Administration 2005). The case is useful because the policy outcome was concrete and the relevant outcome of youth alcohol-related crashes was closely tied to the policy margin.

The main advantage of this case is that the evidence drew on observed policy variation across states and over time. The main limitation is that cross-state comparisons can be affected by contemporaneous changes in enforcement, driving behavior, alcohol regulation, or broader social trends. The credibility of the evidence therefore depends on whether the timing and comparison groups adequately isolate the role of drinking-age laws from other factors.

Nutrition Policy: SNAP under the Obama Administration. The Supplemental Nutrition Assistance Program (SNAP) is a large, ongoing federal program, making it difficult to evaluate through a single experimental design. The evidence base instead draws on multiple observational and quasi-experimental approaches, including state policy variation, the county-level rollout of the Food Stamp Program, changes in benefit generosity, administrative data, and variation in the timing of monthly benefit receipt. These studies examine several outcomes central to nutrition policy: poverty, food insecurity, food consumption, health, academic performance, and long-run human capital (Council of Economic Advisers 2015).

On poverty and food insecurity, SNAP is associated with substantial reductions in material hardship. In 2014, SNAP benefits lifted millions of people above the poverty line, including a large number of children. Studies of food insecurity use several empirical strategies, including comparisons around program entry, state-level policy variation, and changes in benefit levels. The 2009 benefit increase under the American Recovery and Reinvestment Act provides one policy-variation example: higher SNAP benefits were followed by improvements in food security among low-income households, with especially large effects on very low food security.

On food consumption and nutrition, the evidence highlights the importance of program design. SNAP benefits are distributed monthly, and recipient households often spend down benefits before the end of the month. Food expenditures and caloric intake tend to decline over the benefit cycle, with estimates

suggesting sizable reductions in intake late in the month. The American Recovery and Reinvestment Act of 2009 (ARRA) benefit increase is also informative on this margin, since higher benefit levels appear to have smoothed consumption over the month. This evidence connects a specific administrative feature of the program, monthly benefit delivery, to household food purchasing and consumption behavior.

On health and human capital, the evidence covers both short-run and long-run outcomes. For adults, SNAP and food insecurity are linked to self-assessed health, medical care use, and hospitalizations for diet-sensitive conditions such as hypoglycemia among low-income diabetics. For children, the evidence includes birth outcomes, school performance, disciplinary incidents, and adult outcomes. Studies exploiting the staggered county rollout of the Food Stamp Program find that early-life exposure to food assistance improved birth outcomes and was associated with better long-run health, educational attainment, and economic self-sufficiency among disadvantaged children.

Benefit adequacy is a central policy issue in this case. Persistent food insecurity among SNAP households, end-of-month declines in food consumption, and research on the Thrifty Food Plan all suggest that current benefit levels may fall short of the cost of an adequate diet. The concern is not only the dollar amount of benefits, but also the assumptions used to set them: the food basket may not match actual consumption patterns, meal preparation time assumptions may be unrealistic for working families, geographic food-price differences are only partially addressed, and inflation adjustments may lag behind actual food costs.

The policy response under the Obama Administration included several nutrition-program initiatives. ARRA temporarily increased maximum SNAP benefits, raised the minimum benefit, allowed states to suspend certain time limits for unemployed able-bodied adults, and provided additional administrative funding. Other initiatives targeted child nutrition, including changes to the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) food package, the Community Eligibility Provision for free school meals in high-poverty areas, expanded summer meals, and the Summer Electronic Benefits Transfer for Children demonstration. The strength of the case is the breadth of outcomes and evidence reviewed. Its limitation is that the policy connection is diffuse: the evidence supports a broader nutrition-policy agenda rather than identifying a study that directly led to policy change.

5.2 Experimental Evidence

Job Training and Workforce Development: The National JTPA Study. The National Job Training Partnership Act (JTPA) Study (1987–1989) represents a foundational contribution to the empirical evaluation of labor market policies. It is widely regarded as a landmark application of randomized controlled trials to U.S. social policy (Bloom et al. 1997). Prior to this evaluation, assessments of federal job training relied primarily on non-experimental methods, which produced inconsistent findings and raised persistent concerns about selection bias—specifically the “Ashenfelter’s dip” phenomenon, where individuals enter training programs following a transitory decline in earnings, making simple pre-post comparisons misleading.

The JTPA was the primary federal initiative providing services to economically disadvantaged adults and youth. The evaluation randomized more than 20,000 eligible applicants across 16 local service delivery areas into treatment and control groups, with the latter denied access to JTPA services for an

eighteen-month period. This experimental design ensured balance in both observed and unobserved characteristics across groups, setting a new rigorous standard for policy evaluation.

The study's findings revealed substantial heterogeneity in program impacts, challenging the prevailing assumption that training would yield uniform benefits. For adult participants, the program generated meaningful earnings gains, with particularly strong effects for women. Interestingly, these gains were driven primarily by increased labor supply rather than higher hourly wages, suggesting the program helped participants find more consistent work rather than higher-paying roles (Bloom et al. 1997). In contrast, the results for youth were largely null or negative; no statistically significant benefits were found for young women, and adverse earnings effects were observed for certain groups of young men.

Cost-benefit analyses further underscored this divergence. While the program generated positive net social returns for adults, the combination of high program costs and weak impacts for youth resulted in unfavorable cost-benefit profiles for younger participants. These unexpected and consequential results prompted a radical reassessment of federal employment strategies. The JTPA was eventually replaced by the Workforce Investment Act (1998) and later the Workforce Innovation and Opportunity Act (WIOA), which emphasized more intensive service delivery for youth and a shift toward performance-based accountability (Heckman et al. 1999).

Education Production Functions: Tennessee Project STAR. The Student-Teacher Achievement Ratio experiment (Project STAR), conducted between 1985 and 1989, is the foundational randomized trial in education economics (Krueger 1999). The study addressed a long-standing policy debate: whether the substantial fiscal costs of reducing class size were justified by measurable educational benefits. The Tennessee legislature authorized the experiment as a large-scale evaluation, making it a rare example of experimental evidence being generated specifically to inform state-level legislative decisions.

The experiment involved 79 schools and over 11,000 students. Entering kindergarteners were randomly assigned to one of three environments: small classes (13–17 students), regular classes (22–25 students), or regular classes with a full-time teacher's aide. Students remained in their assigned types through third grade to capture cumulative exposure. Initial results demonstrated that students in small classes achieved substantially higher test scores in reading and math. Crucially, the presence of a teacher's aide in regular classrooms did not yield comparable gains, suggesting that the benefits of small classes stem from fundamental changes in teacher-student interaction and instructional dynamics rather than just increased adult supervision (Krueger 1999).

When students transitioned back into regular classrooms in later grades, the achievement advantages diminished—a pattern known as “test score fade-out.” However, subsequent follow-up studies using linked administrative data offered a different perspective. Despite the attenuation of test score differences by middle school, students assigned to small classes exhibited significantly better adult outcomes: higher rates of college attendance, higher earnings at age 27, and lower rates of criminal involvement (Chetty et al. 2011).

These findings played a central role in shaping modern human capital theory. The persistence of labor market gains despite test score fade-out suggests that small class environments promote non-cognitive skills—such as persistence, social behavior, and engagement—that are imperfectly captured by tests but highly valued in the labor market. However, the broader policy relevance of Project STAR

is cautioned by mixed results from subsequent scale-ups. In California, for instance, universal class-size reduction failed to replicate STAR’s magnitude, likely because the rapid hiring of inexperienced teachers diluted instructional quality (Jepsen, Rivkin 2009). This contrast highlights the importance of general equilibrium considerations and implementation capacity when scaling experimental findings.

Teacher Incentives and Personnel Economics: The POINT Experiment. The Project on Incentives in Teaching (POINT) in Nashville (2006–2009) represents a prominent RCT evaluating performance-based pay in education (Springer et al. 2010). Personnel economics predicts that financial incentives tied to performance should increase worker effort. POINT tested whether merit-based bonuses for teachers—ranging up to \$15,000—translated into measurable gains in student math scores.

Across the three-year study, the evaluation found no statistically significant differences in student achievement between classrooms taught by incentivized teachers and those in the control group. While teachers viewed the bonuses favorably, survey evidence indicated that the incentives did not lead them to change their instructional practices. A majority of teachers stated they were already exerting near-maximum effort prior to the introduction of performance pay.

These null results suggest that financial rewards alone may be insufficient when workers are already operating near perceived capacity or when they lack the specific pedagogical tools to improve outcomes further. The experiment redirected policy attention away from narrow performance-pay schemes toward interventions focused on instructional capacity, such as teacher coaching and curriculum reform. POINT serves as a canonical example of how experimental evidence can clarify production constraints and prevent the misallocation of resources toward ineffective policy margins (Springer et al. 2010).

Early Childhood Human Capital: Tennessee Voluntary Pre-K. The evaluation of the Tennessee Voluntary Pre-K (TN VPK) program offers evidence that contrasts with earlier findings from high-intensity preschool interventions like the Perry Preschool Project. Exploiting oversubscription lotteries, the study estimated program impacts under conditions representative of large-scale public provision (Durkin et al. 2022).

A key distinction in this evaluation was the nature of the comparison group. While control children in earlier studies often stayed at home, many children in the TN VPK control group attended other childcare centers or Head Start. Thus, the estimated effects reflect a comparison against prevailing early education options rather than the absence of formal care. While participation in TN VPK was associated with higher school readiness at kindergarten entry, performance converged by third grade. Longer-term follow-up through middle school documented adverse outcomes for program participants, including higher disciplinary incidents and special education placements (Durkin et al. 2022).

These findings have contributed to the “sustaining environments” hypothesis, which posits that early gains may depreciate if subsequent schooling environments are of insufficient quality to build on preschool-acquired skills. The TN VPK evidence cautions against extrapolating results from small-scale, intensive programs to broad state systems characterized by heterogeneous quality. It highlights that the effectiveness of early interventions depends not only on the initial program but also on the institutional continuity across the educational life cycle (Pearman et al. 2020).

The Philadelphia Postcard Study. A field experiment evaluated whether reducing informational barriers could increase participation in a utility payment assistance program. Low-income households were randomized to receive various outreach materials (Teodoro, Smith 2025). The results showed that simple, low-cost black-and-white postcards were just as effective as expensive, elaborate letters in driving enrollment. Most of the gains came from the initial mailing, with diminishing returns to repeated outreach. Similar interventions in schools reduced student absenteeism simply by correcting parental misperceptions about their child’s attendance record (Rogers, Feller 2018). These studies demonstrate that behavioral nudges targeting attention constraints can generate measurable improvements without altering program eligibility or generosity.

The BIAS Project. The Behavioral Interventions to Advance Self-Sufficiency (BIAS) project applied insights to welfare-to-work programs. The study experimentally varied notification letters for TANF recipients regarding required orientations. Results showed that *loss-framed* messages (emphasizing the loss of benefits) significantly increased attendance compared to standard or *gain-framed* notices (Farrell et al. 2016). However, the effects were short-lived; long-term job placement showed little difference across groups. This indicates that while behavioral messaging effectively manages short-run compliance, it does not overcome structural barriers such as skill deficits or transportation constraints.

Social Capital and Youth Development: SMILE. The Study of Mentoring in the Learning Environment (SMILE) evaluated school-based mentoring for at-risk students in San Antonio (Karcher 2008). Students were randomly assigned to receive standard supportive services with or without assignment to a volunteer mentor, isolating the incremental contribution of mentoring. While short-term results were mixed—showing improved self-esteem for elementary boys but disengagement for high school boys—a ten-year administrative follow-up revealed a “sleeper effect.” Individuals assigned to a mentor were substantially less likely to experience arrest and more likely to pursue post-secondary education (Karcher 2020). These effects were most pronounced for female participants. The SMILE evidence illustrates the importance of long-horizon evaluation; by fostering resilience and social capital, mentoring may alter life trajectories in ways that immediate behavioral indicators cannot capture.

TABLE 1. COMPARATIVE SUMMARY OF METHODOLOGICAL APPLICATIONS

Panel A. Observational Evidence			
Case Study	Method	Outcome/Effect	Policy Impact/Takeaway
Smoking and Lung Cancer	Observational evidence review	Disease risk	Authoritative synthesis of epidemiological and biomedical evidence helped support warning labels and advertising restrictions (U.S. Department of Health, Education, and Welfare 1964; Centers for Disease Control and Prevention 2024).
Minimum Legal Drinking Age	Observational / policy variation	Crash risk	Cross-state evidence on alcohol-related crashes among young drivers helped support the move toward a uniform minimum legal drinking age of 21 (National Highway Traffic Safety Administration 2005).
SNAP under the Obama Administration	Observational / quasi-experimental evidence review	Food security and long-run outcomes	CEA synthesis of SNAP evidence supported the policy rationale for benefit expansions, summer food assistance, WIC improvements, and expanded school meals (Council of Economic Advisers 2015).
Panel B. Experimental Evidence			
Case Study	Method	Estimand	Policy Impact/Takeaway
JTPA Training	RCT	ITT/ATT	Pivot toward comprehensive, intensive youth services (Bloom et al. 1997).
Project STAR	RCT	ATE	Importance of non-cognitive skills; supply-side scale-up risks (Chetty et al. 2011).
POINT Incentives	RCT	ATE	Limits of pay-for-performance; shift to teacher capacity building (Springer et al. 2010).
TN-VPK	RCT	ITT	Importance of K–3 alignment to prevent achievement fade-out (Durkin et al. 2022).
Phila. Postcard	Field RCT	ATE	Low-cost information nudges can effectively reduce “sludge” (Teodoro, Smith 2025).
TANF Letters	Beh. RCT	ATE	Loss-aversion increases compliance but not long-term employment (Farrell et al. 2016).
SMILE Mentoring	RCT	ITT	Long-term reduction in crime despite mixed short-term results (Karcher 2020).

6 Challenges and Critiques of Evidence-Based Policy

Evidence-based policymaking is not merely a technical exercise in producing valid estimates. It is a process through which empirical claims are generated, interpreted, communicated, and implemented

within political and institutional environments. At each stage, evidence can inform policy, but it can also be distorted, overinterpreted, selectively used, or rendered irrelevant by constraints outside the research design itself. The central critique of EBP, therefore, is not that evidence is unimportant. Rather, it is that evidence rarely speaks with the level of precision, generality, and finality that policy institutions often demand.

This section reviews several recurring challenges. Some arise from the production of evidence, including ethical limits to experimentation, untestable assumptions in observational studies, and threats to internal validity. Others arise from the translation of evidence into policy, including external validity, scaling, implementation fidelity, and political selection in the use of research. A final set of challenges concerns the data infrastructure on which modern EBP increasingly depends, including measurement error, privacy, and the governance of administrative data.

6.1 The Demand for Certainty in an Uncertain World

A primary challenge to EBP is the mismatch between the norms of scientific research and the demands of the political arena. Rigorous empirical research typically produces probabilistic statements: an estimated effect, a confidence interval, a set of identifying assumptions, and a discussion of limitations. Policymakers, by contrast, often face immediate decisions under political, fiscal, and administrative pressure. They may need to decide whether to fund a program, expand an intervention, or terminate an existing policy before a complete body of evidence is available.

This mismatch creates what Manski (2013) calls “incredible certitude”: the tendency to present uncertain evidence as if it delivered definitive answers. The problem is not simply that policymakers demand certainty. Researchers, consultants, advocates, and government agencies may also have incentives to communicate findings in overly confident language. A study that estimates a positive average treatment effect may be summarized as showing that a program “works”, even when the estimate is imprecise, context-specific, or dependent on strong assumptions. Conversely, a statistically insignificant estimate may be interpreted as evidence that a program “does not work”, even when the study is underpowered or the relevant effects are heterogeneous.

The political use of evidence further complicates this process. Evidence is often introduced into policy debates after major positions have already been formed. In such settings, research may be used less as a neutral input into decision-making than as a rhetorical resource to defend existing preferences. Policymakers may emphasize studies consistent with their ideological commitments, discount studies that conflict with their priorities, or demand a standard of proof from disfavored policies that they do not require from favored ones (Cairney 2016). EBP therefore does not eliminate politics; at best, it disciplines political debate by making assumptions, trade-offs, and empirical claims more explicit.

6.2 Internal Validity, Ethics, and the Limits of Randomization

Randomized controlled trials are often treated as the most credible design for estimating causal effects because random assignment can eliminate selection bias. Yet randomization also raises ethical and practical concerns. In many policy domains, treatment cannot be assigned purely for research purposes. Denying services to vulnerable populations, delaying access to benefits, or varying exposure to poten-

tially harmful interventions may be politically or morally unacceptable (Ravallion 2020). Even when randomization is feasible, the act of experimentation itself may alter the policy environment by changing expectations, implementation effort, or participant behavior.

EBP has developed several design responses to these concerns. Randomized phase-in designs can be used when a program will eventually reach all eligible participants but resources require gradual rollout. Lotteries can serve as transparent rationing mechanisms when demand exceeds available capacity. Encouragement designs can randomize information or incentives without directly denying access to services. These approaches reduce some ethical concerns, but they do not eliminate the deeper tension between knowledge generation and immediate welfare (Glennerster, Takavarasha 2013).

Moreover, even well-designed RCTs may face threats to internal validity. Noncompliance can create a gap between assignment and actual treatment receipt. Attrition can undermine baseline balance if treated and control units leave the study at different rates. Spillovers can contaminate the control group when untreated individuals are indirectly affected by treated individuals, particularly in labor markets, schools, public health, or networked communities. These problems do not make RCTs uninformative, but they mean that the relevant estimand may differ from the simple average treatment effect imagined in a textbook experiment. In many policy settings, the intention-to-treat effect, local average treatment effect, or equilibrium effect is more relevant than the effect of treatment under full compliance and no interference.

6.3 External Validity, Scaling, and Implementation Fidelity

Even when a study is internally valid, its relevance for policy depends on whether the estimated effect generalizes beyond the original context. A central critique of EBP is that evidence generated in one population, time period, or institutional setting may not travel well to another. The effectiveness of a job training program may depend on local labor demand, employer networks, and participant selection. The effect of an educational intervention may depend on teacher quality, school leadership, or parental responses. The impact of a health intervention may depend on baseline access to care and administrative capacity.

This challenge becomes especially acute when programs are scaled. Small pilots often operate under favorable conditions: researchers are closely involved, implementers are well trained, participants may be unusually motivated, and monitoring is intensive. When the same program is transferred to routine government delivery, these conditions may disappear. The resulting “voltage drop” describes the deterioration of impacts or benefit-cost ratios as interventions move from controlled pilot

central government may adopt an evidence-based program, but its realized impact depends on how local implementers interpret eligibility rules, allocate effort, communicate with beneficiaries, and respond to monitoring. This principal-agent problem implies that the success of EBP depends not only on the credibility of the original research but also on the institutional architecture through which the policy is delivered.

6.4 Observational Evidence and the Problem of Unobserved Counterfactuals

Observational studies occupy an ambiguous position in EBP. On the one hand, they are indispensable. Many important policies cannot be randomized, either because they are already implemented at scale, because randomization would be unethical, or because the relevant outcomes unfold over long horizons. Observational studies using administrative data, longitudinal surveys, or naturally occurring policy variation are often the only feasible way to evaluate large social programs, tax policies, labor regulations, education reforms, and health systems. They can also be faster and less expensive than new experiments because the relevant data may already exist.

On the other hand, observational studies rely on assumptions to recover missing counterfactual outcomes. Some assumptions, such as overlap or pre-treatment covariate balance, can be partially examined. Others, such as the absence of unobserved confounding, exclusion restrictions, or parallel trends, are inherently untestable. This does not mean that all observational studies are merely “leaps of faith”; rather, it means that their credibility depends on the strength of the research design, the institutional source of variation, the transparency of the identifying assumptions, and the plausibility of alternative explanations.

The central risk is that observational evidence can be mistaken for causal evidence when the design does not justify such an interpretation. Selection bias, reverse causality, time-varying confounding, anticipation effects, and measurement error can all produce associations that resemble policy impacts. Even sophisticated econometric methods cannot fully compensate for weak research design. Matching, weighting, fixed effects, difference-in-differences, instrumental variables, and machine-learning adjustment each address particular threats, but none eliminates the need for a credible counterfactual argument.

A balanced view is therefore necessary. Observational studies should not be dismissed simply because they are non-randomized. Some observational designs, especially those exploiting sharp institutional rules or rich longitudinal data, can be highly credible. At the same time, their assumptions should be stated clearly, and sensitivity analyses should be used to quantify how strong violations would need to be to alter the conclusions. For policymaking, the question is not whether a study is experimental or observational in a mechanical hierarchy, but whether the

analyses, and then used to justify policy action. The relevant evidentiary object is therefore not always a single study, but a chain linking research findings to policy-facing documents and, ultimately, to public decisions.

6.5 From Observational Evidence to Policy: Citation Chains

One way to understand the policy influence of observational research is through what may be called a “citation chain”. In such cases, observational studies or evidence reviews are first synthesized in an official report, regulatory impact analysis, advisory recommendation, court record, or legislative document. These policy-facing documents then provide part of the rationale for public action. The chain is rarely as simple as one study causing one law. Rather, observational evidence becomes policy-relevant when it is accumulated, assessed, and translated by institutions that policymakers regard as authoritative.

The strongest examples often come from public health, environmental regulation, and transportation safety. These are domains in which randomization is frequently infeasible or unethical, and where long-run harms must be inferred from cohort studies, case-control studies, cross-jurisdictional variation, toxicological evidence, and systematic reviews. The tobacco control and minimum legal drinking age examples discussed previously provide a canonical examples. The credibility of the evidence came not from a single research design but from the convergence of many observational, clinical, and biological sources.

An additional example is environmental regulation. The regulation of fine particulate matter relies heavily on epidemiological evidence linking long-term pollution exposure to mortality and morbidity. Regulatory impact analyses for particulate-matter standards use concentration-response functions derived from cohort-based mortality studies and related epidemiological evidence to estimate health benefits and monetize avoided premature deaths (Pope et al. 1995; Pope et al. 2002; U.S. Environmental Protection Agency 2024b). In this setting, observational evidence is not merely background information. It becomes an input into the formal benefit-cost calculations that help justify regulatory standards. The regulation of lead under the National Ambient Air Quality Standards follows a related process. Scientific evidence is synthesized through air-quality criteria documents, Integrated Science Assessments, risk and exposure assessments, and policy assessments, which together form the scientific basis for decisions about whether existing standards adequately protect public health (U.S. Environmental Protection Agency 2024a; U.S. Environmental Protection Agency 2025).

experimentation. In many policy domains, the relevant counterfactual cannot be produced by random assignment. It would be unethical to randomize smoking, air pollution exposure, lead exposure, infant sleep position, or drunk-driving risk. In such cases, credible evidence accumulates through triangulation: repeated observational findings, biological or behavioral mechanisms, robustness across populations, and formal synthesis in expert assessments. The policy influence of observational evidence therefore often depends on whether it survives review by institutions designed to translate science into action.

At the same time, citation chains do not eliminate the usual limitations of observational evidence. The fact that a study is cited in a government report does not make its identifying assumptions true. Cohort studies may suffer from residual confounding; cross-state comparisons may be affected by contemporaneous policy changes; administrative records may measure policy exposure or outcomes imperfectly. Moreover, the policy translation process can selectively emphasize some studies over others. For this reason, the relevant standard is not whether evidence is experimental or observational, but whether the evidence base is sufficiently credible, transparent, and cumulative for the policy question being asked.

6.6 Evidence Use, Political Selection, and Institutional Incentives

Even when credible evidence exists, it may not be used. Political institutions do not passively absorb research findings; they filter evidence through electoral incentives, budget constraints, interest-group pressure, administrative capacity, and public opinion. A policy with strong evidence may fail because its benefits are diffuse and long-run while its costs are concentrated and immediate. Conversely, a policy with weak evidence may persist because it serves a politically influential constituency or because terminating it would create visible losses.

This creates an important asymmetry in the politics of evidence. Evidence of program failure does not automatically lead to program termination, and evidence of success does not automatically produce scale-up. Policymakers must consider not only average effectiveness but also who benefits, who pays, who is organized, and when the effects become visible. In labor, education, and population policy, many of the most important outcomes—earnings, health, fertility, mobility, or intergenerational transmission—emerge over years or decades. Political time horizons are often much shorter.

error (Card et al. 2010). These data have greatly expanded the scope of policy evaluation, especially in domains where outcomes are difficult to measure through short-term surveys.

Yet administrative data introduce their own limitations. They are collected for program administration, not necessarily for research. Variables may reflect eligibility rules, reporting incentives, bureaucratic classifications, or strategic behavior by agencies and participants. A recorded outcome may differ from the underlying concept of interest. For example, benefit receipt is not the same as need, reported earnings may omit informal work, and disciplinary records may reflect both student behavior and school enforcement practices. As EBP increasingly incorporates machine learning, text, images, and linked records, measurement error and construct validity become central concerns rather than secondary technical details.

Data access also creates a governance problem. The same linkages that make administrative data valuable for research can increase privacy risks. Traditional de-identification is increasingly fragile in an era when multiple datasets can be combined to re-identify individuals (Sweeney 2002). Differential privacy and other formal privacy protections attempt to address this risk by introducing statistical noise or access restrictions, but these protections involve trade-offs. Stronger privacy guarantees may reduce precision, limit small-area analysis, or complicate replication (Abowd 2018). Evidence-based policymaking therefore requires not only better data but also legitimate rules governing access, privacy, transparency, and public trust.

These challenges suggest a more modest but more defensible vision of EBP. Evidence can improve policy by clarifying causal relationships, documenting trade-offs, and revealing unintended consequences. But evidence cannot eliminate uncertainty, resolve normative conflict, or guarantee faithful implementation. The value of EBP lies not in replacing politics with technocracy, but in making policy choices more disciplined, transparent, and accountable.

7 Summary

This chapter provides a high-level overview of evidence-based policymaking (EBP), emphasizing that it is not a singular algorithmic procedure but rather an iterative process that must confront and balance the fundamentally different objectives of diverse

validity to the demand for “incredible certitude” to the “voltage effect” often observed when scaling interventions, are discussed.

Looking toward the future, the scope of EBP is expanding through the use of high-frequency data and predictive analytics. Technological partnerships now enable the real-time tracking of economic activity, as evidenced by the *Opportunity Insights Economic Tracker*, which synthesized payroll and credit card data to provide weekly insights into the COVID-19 recovery (Chetty et al. 2024). Furthermore, the rise of machine learning and policy learning algorithms offers the potential to design optimal treatment rules that maximize welfare subject to budget and equity constraints.

Ultimately, evidence-based policymaking is not a panacea for political conflict, nor is it a replacement for the normative judgments that must guide a society. It is, instead, a disciplined framework for navigating the 21st century’s complex policy landscape. By grounding decisions in the potential outcomes framework—while remaining clear-eyed about the institutional and ethical frictions of implementation—economists can ensure that the link between evidence and action results in more effective, equitable, and resilient public interventions.

8 Cross References

- Causality
- Randomized Control Trials for Policy Evaluation
- Causal Analysis of Social Policies

9 Acknowledgments

Responsible Section Editor: Enrico Rettore. The article has benefited from valuable comments of the editors and anonymous referees. There is no conflict of interest.

Index

behavioral economics, 3

Causal Forests, 17

Cost-Benefit Analysis, 18

counterfactual, 7

Difference-in-Differences, 15

Double/Debiased Machine Learning, 11

Equity, 18

Evidence-based policymaking, 1

external validity, 10

externalities, 3

general equilibrium, 10

identification, 8

impact evaluation, 7

information asymmetries, 3

Instrumental variables, 15

internal validity, 10

Machine learning, 16

nudges, 4

Policy analysis, 7

Policy Learning, 17

potential outcomes, 8

propensity score, 11

Quasi-experimental methods, 1

Randomized controlled trials, 1

Regression Discontinuity, 16

spillovers, 10

Stable Unit Treatment Value Assumption, 8

Synthetic Control Methods, 15

treatment effects, 1

References

- Abadie A (2021) Using synthetic controls: Feasibility, data requirements, and methodological aspects. *J Econ Lit* 59(2):391–425.
- Abowd JM (2018) The U.S. Census Bureau adopts differential privacy. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18. Association for Computing Machinery, London, United Kingdom, p 2867.
- Akerlof GA (1978) The market for “lemons”: Quality uncertainty and the market mechanism. In: *Uncertainty in Economics*. Elsevier, p 235–251.
- American Academy of Pediatrics Task Force on Infant Positioning and SIDS (1992) Positioning and SIDS. *Pediatrics* 89(6):1120–1126.
- Anderson ML (2008) Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *J Am Stat Assoc* 103(484):1481–1495.
- Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *J Am Stat Assoc* 91(434):444–455.
- Athey S, Imbens GW (2019) Machine learning methods that economists should know about. *Annu Rev Econ* 11(1):685–725.
- Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.
- Austin PC (2009) Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-Score Matched Samples. *Statistics in Medicine* 28(25):3083–3107.
- A

- Cappelli PH (2015) Skill gaps, skill shortages, and skill mismatches: Evidence and arguments for the United States. *ILR Review* 68(2):251–290.
- Card D, Kluve J, Weber A (2018) What works? A meta analysis of recent active labor market program evaluations. *J Eur Econ Assoc* 16(3):894–931.
- Card DE, Chetty R, Feldstein MS, Saez E (2010) Expanding access to administrative data for research in the United States. SSRN.
- Carneiro P, Heckman JJ, Vytlacil E (2010) Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica* 78(1):377–394.
- Cawley J, Susskind A, Willage B (2020) The impact of information disclosure on consumer behavior: Evidence from a randomized field experiment of calorie labels on restaurant menus. *J Policy Anal Manage* 39(4):1020–1042.
- Centers for Disease Control and Prevention (2024) *A History of the Surgeon General's Reports on Smoking and Health*. <https://www.cdc.gov/tobacco-surgeon-general-reports/about/history.html>. Accessed June 10, 2026.
- Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018a) Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econom J* 21(1):C1–C68.
- (2018b) Double/debiased machine learning for treatment and structural parameters. *Econom J* 21(1):C1–C68.

- Dehejia RH, Wahba S (1999) Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *J Am Stat Assoc* 94(448):1053–1062.
- (2002) Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Rev Econ Stat* 84(1):151–161.
- Drucker L, Mazirov K, Neumark D (2021) Who pays for and who benefits from minimum wage increases? Evidence from Israeli tax data on business owners and workers. *J Pub Econ* 199:104423.
- Dube A (2019) Minimum wages and the distribution of family incomes. *Am Econ J Appl Econ* 11(4):268–304.
- Duflo E, Glennerster R, Kremer M (2007) Using randomization in development economics research: A toolkit. *Handbook of Development Economics* 4:3895–3962.
- Durkin K, Lipsey MW, Farran DC, Wiesen SE (2022) Effects of a statewide pre-kindergarten program on children’s achievement and behavior through sixth grade. *Dev Psychol* 58(3):470–484.
- Einav L, Finkelstein A (2011) Selection in insurance markets: Theory and empirics in pictures. *J Econ Perspect* 25(1):115–138.
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (2009) *Infant Sleep Position and SIDS: Questions and Answers for Health Care Providers*.

- Heckman JJ, Ichimura H, Todd PE (1998b) Matching as an Econometric Evaluation Estimator. *Rev Econ Stud* 65(2):261–294.
- Heckman JJ, LaLonde RJ, Smith JA (1999) The economics and econometrics of active labor market programs. In: *Handbook of Labor Economics*. Elsevier, p 1865–2097.
- Heckman JJ, Moon SH, Pinto R, Savelyev PA, Yavitz A (2010) The rate of return to the HighScope Perry preschool program. *J Pub Econ* 94(1-2):114–128.
- Heckman JJ, Vytlacil E (2005) Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73(3):669–738.
- Hendren N, Sprung-Keyser B (2020) A unified welfare analysis of government policies. *Q J Econ* 135(3):1209–1318.
- Hern'án MA, Robins JM (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 183(8):758–764.
- Hern'án MA, Sauer BC, Hern'andez-D'íaz S, Platt R, Shrier I (2016) Specifying a Target Trial Prevents Immortal Time Bias and Other Self-Inflicted Injuries in Observational Analyses. *Journal of Clinical Epidemiology* 79:70–75.
- Holland PW (1986) Statistics and causal inference. *J Am*

- LaLonde RJ (1986) Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *Am Econ Rev* 76(4):604–620.
- Lechner M (2023) Causal Machine Learning and its use for public policy. *Swiss Journal of Economics and Statistics* 159(1):8.
- Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. *J Econ Lit* 48(2):281–355.
- List JA (2022) The voltage effect: How to make good ideas great and great ideas scale. Currency, New York.
- Madrian BC, Shea DF (2001) The power of suggestion: Inertia in 401(k) participation and savings behavior. *Q J Econ* 116(4):1149–1187.
- Manski CF (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–1246.
- (2013) *Public policy in an uncertain world: Analysis and decisions*. Harvard University Press.
- Martins PS, Jin JY (2010) Firm-level social returns to education. *J Popul Econ* 23(2):539–558.
- Medina PC (2021) Side effects of nudging: Evidence from a randomized intervention in the credit card market. *Rev Financ Stud* 34(5):2580–2607.
- Melly

- Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW (1995) Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults. *American Journal of Respiratory and Critical Care Medicine* 151(3):669–674.
- Pyatt G (1966) Human capital: A theoretical and empirical analysis, with special reference to education the residual factor and economic growth econometric models of education. *Econ J* 76(303):635–638.
- Ravallion M (2020) Should the Randomistas (Continue to) rule? In: Bédécarrats F, Guérin I, Roubaud F (eds.) *Randomized Control Trials in the Field of Development: A Critical Perspective*. Oxford University Press, Oxford, p 47–62.
- Richards GW (2017) How research–policy partnerships can benefit government: A win–win for evidence-based policy-making. *Can Public Policy* 43(2):165–170.
- Robinson PM (1988) Root-N-Consistent Semiparametric Regression. *Econometrica* 56(4):931–954.
- Rogers T, Feller A (2018) Reducing student absences at scale by targeting parents’ misbeliefs. *Nat Hum Behav* 2(5):335–342.
- Rosenbaum PR

- Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J Econometrics* 225(2):175–199.
- Sweeney L (2002) k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5):557–570.
- Teodoro MP, Smith J (2025) Medium and message by mail: A field experiment to promote low-income assistance programs. *Publ Adm Review*. Forthcoming.
- Thaler RH, Benartzi S (2004) Save more tomorrow: Using behavioral economics to increase employee saving. *J Polit Econ* 112(S1):S164–S187.
- Thaler RH, Sunstein CR (2009) *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- U.S. Department of Health, Education, and Welfare (1964) *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Public Health Service Publication No. 1103. Washington, DC: U.S. Public Health Service.
- U.S. Environmental Protection Agency (2024a) *Integrated Science Assessment for Lead*. Final report. Washington, DC: U.S. Environmental Protection Agency.
- (2024b) *Regulatory Impact Analysis for the Final Reconsideration of the National Ambient Air Quality Standards for Particulate Matter*.