ISSN: 1813-4890

A literature review on policy evaluation under causal effects

Houwu Wu

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Anhui 230000, China.

320230027@aufe.end.cn

Abstract

Policy evaluation aims to comprehensively understand the effects of specific policy measures on target variables through an in-depth examination of causal effects. Since the 20th century, causal inference has garnered widespread attention and development in policy evaluation, resulting in numerous empirical research findings and review articles. However, existing reviews in policy evaluation research predominantly focus on specific domains, with relatively limited introductions to policy evaluation methods. This paper provides a comprehensive review of various methods widely used in causal effect studies, with a particular emphasis on their applications in the field of policy evaluation. Firstly, by reviewing and summarizing recent empirical research literature on policy evaluation, the paper revisits traditional empirical evaluation methods, particularly randomized controlled trials. Despite the acknowledged internal validity of this method, it faces challenges related to external validity and implementation. Secondly, the paper concentrates on emerging causal inference methods in recent years, including regression discontinuity design, propensity score matching, and difference-in-differences. These methods aim to overcome limitations of traditional approaches and exhibit notable performance in addressing causal issues within real-world policy contexts. Lastly, the paper underscores the significance of policy evaluation and anticipates future trends, encompassing aspects such as data structures, causal inference methods, and the integration of machine learning with big data. This comprehensive review enhances the understanding of policy evaluation methods and provides valuable insights for future research.

Keywords

Causal Effects, Policy Evaluation, Theoretical Methods.

1. Introduction

Evaluating the effectiveness of specific policies in a scientific, objective, and accurate manner is crucial for determining whether these policies can effectively intervene in the economy and achieve their intended goals. Such evaluations are essential for providing the government with the basis for decision adjustments, deepening, and optimization. In this broad and complex research field, the concept of causal effects has become increasingly important. Causal effects involve the impact that a specific policy or intervention has on observed changes. This impact needs not only to be quantified but also to exclude other potential explanatory factors. An in-depth study of causal effects is precisely the way we can more accurately understand the relationship between policy implementation and social change, helping to avoid the misleading influence of simple correlations on our understanding of policy effects. The core task of policy evaluation is to answer a series of key questions, including examining whether "policy intervention has caused a causal link to policy outcomes" and whether "policy intervention can effectively achieve the expected effects."

ISSN: 1813-4890

For a long time, quantifying the effects of policy evaluations has been a significant challenge. Until the mid-20th century, the introduction of randomized controlled trials (RCTs) provided a solution to this challenge. By randomly assigning experimental and control groups, researchers could more effectively control for other variables, thereby more accurately identifying causal effects [1, 2]. However, given that policy effect studies tend to rely more on natural experiments, the non-random assignment of study subjects inevitably leads to "selection bias" in the samples. This makes it difficult to establish causal links between policy interventions and outcomes, necessitating the development of causal inference methods based on observational studies. In the latter half of the 20th century, with the rise of econometrics, policy evaluation became more quantitative and model-based [3, 4]. Econometric methods enabled researchers to use statistical techniques to control for confounding factors and reduce the issue of selection bias. Methods for causal inference in policy evaluation and their applications are rapidly advancing towards being more positive, refined, and in-depth, and have become important tools for conducting evidence-based research in the field of policy evaluation.

Based on this, this paper organizes and reviews common methods for policy evaluation under causal effects, such as regression discontinuity, propensity scores, and difference-in-differences. The goal of this paper is to provide a comprehensive review of the key advancements in the field of policy evaluation under causal effects, not only discussing its theoretical framework but also examining its application in actual policy formulation and practice. Through this review, we aim to provide a comprehensive and in-depth perspective on policy evaluation research, offering valuable insights for future research and policy decision-making.

2. Empirical Analysis of Policy Evaluation

Since the mid-20th century, causal inference has increasingly garnered attention from policy evaluators. Prior to this, most policy evaluators employed traditional research methods, such as surveys and interviews, to indirectly infer policy effects by analyzing the differences in data before and after policy implementation. Causal inference techniques provide a quantitative evaluation of policy implementation, thereby revealing policy effects from another perspective. In the field of policy evaluation, both domestic and international scholars have dedicated themselves to the in-depth study of causal effects, continuously improving empirical evaluation methods and striving for more accurate and comprehensive evaluation approaches to better meet practical needs. By enhancing the accuracy of evaluation methods, scholars have successfully provided crucial information to governments and organizations, offering strong references for policy formulation, implementation, and adjustment. The application of causal effects has increased the effectiveness of policy evaluations, playing an increasingly important role in meeting societal needs and addressing issues.

Under causal inference, significant progress has been made in policy evaluation in terms of analytical methods and the integration with policy practice (Cox,1958; Rubin,1974; Heckman,1979; Angrist et al,1996; Pearl,2000; Dehejia and Wahba,2002)[5-10]. First, the methods for evaluating policy effects have been continuously optimized and improved. In the estimation of causal effects, a variety of methods have been widely adopted for policy evaluation, including propensity scores (Heckman et al,1996)[11], regression discontinuity (Thistlethwaite and Campbell,1960)[12], and difference-in-differences methods (Wooldridge and Imbens,2007)[13], all of which have shown significant application in empirical research. Meanwhile, algorithmic analysis has also been continuously refined and improved, introducing methods such as the generalized propensity score (Feng et al,2012)[14], quantile regression discontinuity (Imbens and Lemieux,2008)[15], and double robust estimation (Funk et al,2011)[16], effectively enhancing the original methods. Second, in terms of hot topics in policy evaluation research, the focus has been on areas such as single-policy binary treatment effects,

ISSN: 1813-4890

single-policy multi-valued treatment effects, and multiple-policy treatment effects. Although substantial empirical research has been accumulated in the first two areas, research on multiple-policy treatment effects remains relatively limited. This further underscores the importance of in-depth research into multiple-policy treatment effects to enrich and expand the content of policy evaluation research, better serving practical needs. Single-policy binary treatment methods involve binary processing of policy implementation (i.e., presence or absence), followed by comparing the differences between the treatment group (the group receiving the policy) and the control group (the group not receiving the policy) to infer the estimated policy effects (Angrist et al,1996; Acemoglu and Angrist,2000)[8, 17].

Single-policy multi-valued treatment methods involve examining a single policy where researchers consider multiple treatment options, such as different levels or forms of policy acceptance, and perform multi-valued processing. Subsequently, these treatment groups are compared to evaluate the policy effects (Angrist and Evans, 1996; Sneyers and Witte, 2018)[18, 19]. Multiple-policy treatment refers to researchers examining and comparing the impact of various different policies or combinations of policies on specific issues or groups. This involves simultaneously handling multiple policy variables to comprehensively assess their effects (Greenwald et al., 1996)[20]. For detailed information, refer to Table 1.

Table 1 Hot Topics and Representative Literature in Policy Evaluation Research

Research Direction	Research Domain	Representative Literature	Main Research Conclusions
Single-policy binary treatment	Macroeconomics	Dell, 2010	The study explores the impact of the Mita labor system implemented by Spain in certain regions of Peru from the 16th to the 19th century on economic development [21].
	Medium-scale	Chen et al., 2013	Using a geographic regression discontinuity design, the study investigates the impact of the centralized winter heating system in northern regions of China on the average life expectancy of residents [22].
	Microeconomics	Vandenberghe and Robin, 2004	Using data from the 2000 OECD PISA survey, the study employs instrumental variables and Heckman's two-stage method to investigate the impact of private versus public education on student performance [23].
Single-policy multi-valued treatment	Education	Qi Zhanyong and Du Yue, 2023	By integrating mainstream quantitative analysis methods to evaluate the impact of educational policy implementation, the study proposes an actionable logic from a feasible methodological perspective [24].
	Healthcare	Yang et al., 2016	Using a multi-valued treatment effects model with three treatment choices, the study conducts a comparative analysis of the efficacy

ISSN: 1813-4890

			of two pyrethroid-based medications for treating fibromyalgia [25].
	Labor and Employment	Zhang Peiying et al., 2019	Utilizing the generalized propensity score matching method, the study analyzes the marginal impact of the number of children born on women's wage income [Zhang Peiying 2019 The Marginal Impact of Fertility on Women's Wage Income].
Multiple-policy treatment	Environment	Chabé-Ferret, 2013	The study examines the impact of various agri-environmental policies in France on crop farming [26].

To understand the current landscape of policy evaluation under causal effects, we searched for literature published before 2023 in CNKI and Web of Science using the keywords: Causal Effects, Policy Evaluation, Policy Effects, Causal Inference, Policy Impact Assessment, Policy Studies Review, and Government Policy Evaluation, as shown in Table 2. As previously mentioned, although the literature on policy evaluation under causal effects is relatively abundant, literature reviews on policy studies are comparatively scarce. Therefore, this paper emphasizes the necessity of conducting a literature review based on policy evaluation under causal effects. By organizing and summarizing the commonly used methods in the field of causal inference for policy evaluation, this paper aims to provide profound insights for policy evaluators.

Table 2 Keyword Search Results for Policy Evaluation under Causal Effects

Keyword	Search Engine	
	CNKI	Web of Science
Causal Effects	835	1454
Policy Evaluation	3877	3630
Policy Effects	1579	2026
Causal Inference	3888	3792
Policy Impact Assessment	30	498
Policy Studies Review	0	35
Government Policy Evaluation	0	56

3. Theoretical Methods for Policy Evaluation

To understand the effectiveness of policy implementation, it is essential to evaluate the causal effects before and after the policy is enacted. Causal effects refer to changes in outcomes that are attributable to specific factors. In policy evaluation, we focus on the impact that policies or interventions have on observed changes. A positive causal effect indicates that the policy has led to beneficial changes, while a negative causal effect suggests that the policy has caused adverse outcomes. Identifying causal effects necessitates the exclusion of other potential explanatory factors, allowing us to infer with greater confidence that the observed changes are causally related to the policy implementation. This deep understanding is crucial for accurately assessing policy effectiveness and providing insightful guidance for future policy-making.

Statisticians formalize causal effects using the potential outcomes model. Neyman provided a mathematical definition of causal effects in experimental research [27], defining the Average Treatment Effect (ATE) as the average change induced by the policy across the entire study sample. The ATE can be calculated by comparing the average differences between the treatment group and the control group before and after policy implementation.

ISSN: 1813-4890

$$ATE = E\left[Y_i \mid T = 1\right] - E\left[Y_i \mid T = 0\right] \quad (1)$$

In this context, Y_i is the observed outcome variable, T is the treatment group indicator variable (1 for the experimental group and 0 for the control group), and E represents the expected value. In the evolution of academic research, the potential outcomes framework proposed by scholars such as Rubin (1973, 1974, 1977, 1978) has become the dominant analytical framework for observational studies [28, 6, 29, 30]. Causal effects are defined as the difference between potential outcomes, and strategies for identifying and estimating causal effects have been developed within the counterfactual or potential outcomes framework. While controlling for confounding factors is the core method for studying causal relationships in classical econometrics, it is often quite challenging to fully control all confounding factors in practical problems. Therefore, the current mainstream approach in causal effect research is to overcome endogeneity issues through innovative observational study designs, providing more reliable and comprehensive references for policy-making.

3.1. Randomized Controlled Trials

Randomized Controlled Trials (RCTs) have emerged as the new gold standard in the fields of development economics and institutional analysis (Webber and Prouse, 2018) [31]. The core principle of RCTs lies in the random allocation of study subjects into experimental and control groups to ensure that characteristics other than the policy impact are randomly distributed. The experimental group receives a specific policy or intervention, while the control group remains unaffected, thereby creating a relatively independent benchmark in contrast to other factors. The introduction of randomness eliminates inherent selection bias, enhancing the internal validity of the research.

RCTs have widespread applications in policy evaluation. In the education sector, RCTs are extensively used to thoroughly assess the effects of various educational policies (Morrison, 2001; Torgerson and Torgerson, 2012; Connolly, 2018) [32-34]. This research method provides a highly scientific and objective evaluation framework. By introducing randomness into the experimental and control groups, it helps eliminate potential confounding factors, thereby more accurately measuring the impact of educational policies on students, educational institutions, and the overall education system. In the evaluation of health policies (Dobbins et al., 2009; Murphy et al., 2012; Bonell, 2012), RCTs are widely applied to assess the effects of various medical policies, health interventions, and drug treatments [35-37]. By randomly assigning patients or healthcare institutions, this method helps to more accurately determine the causal effects of treatments. This scientifically rigorous research design provides more credible evidence for medical decision-making and helps to deeply understand the actual impact of specific policies or interventions in promoting health and treatment outcomes. In evaluating employment and training policies (Bond et al., 2018; Alzúa et al., 2013), RCTs can be used to clarify the effects of specific training programs or employment assistance measures on the employment and career development of participants [38, 39]. By adopting a randomized grouping method, this research design helps to more precisely determine the actual effects of training programs or employment assistance in enhancing participants' employment opportunities and career growth. This scientifically rigorous evaluation method not only provides reliable evidence for policy formulation but also promotes a deeper understanding of the effectiveness of employment and training policies. In the evaluation of fiscal and tax policies (Wall et al., 2006; Ni Mhurchu et al., 2010), RCTs can be used to study the impacts of different policies on enterprises, individuals, or the overall economy [40, 41]. By introducing randomness, it is possible to more accurately simulate the complex economic systems in the

ISSN: 1813-4890

real world, while avoiding some of the endogeneity issues that may exist in traditional research methods. In environmental policy evaluation (Bilotta and Milner, 2014; Frederiks et al., 2016; Clasen et al., 2020), RCTs can be employed to deeply explore the impacts of different environmental policies on environmental protection and sustainable development goals [42-44]. By introducing randomness, this method helps to more precisely simulate complex environmental systems and avoid potential endogeneity problems in traditional research methods.

In the field of policy evaluation, RCTs demonstrate unparalleled advantages by randomly assigning treatment and control groups to exclude confounding factors, ensuring the internal and external validity of the results. However, this method also faces several limitations and challenges. RCTs are constrained by individual selection, which may lead to discrepancies between actual effects and provided causal effects. In policy evaluation, obtaining RCT data is often influenced by research costs, ethical constraints, and implementation difficulties. Simultaneously, samples typically exist in a non-random sampling format, leading to non-random allocation of individuals into the treatment group, which may result in significant differences in observable and unobservable factors between groups and trigger endogeneity issues of binary treatment variables. To address this issue, observational studies need to develop credible methods for evaluating causal effects. Using randomized experiments as a benchmark, research design can simulate the allocation mechanism of RCTs by setting causal effect identification assumptions to evaluate policy effects. This method provides a practical and effective approach for causal inference in the absence of RCT data, offering more reliable references for policy decision-making.

In actual policy evaluation, most studies are conducted under non-random trials and observational data analyses. The lack of randomness may lead to significant differences in observable variables between the treatment and control groups, resulting in selection bias and confounding issues. The following sections will focus on methods for inferring causal effects in observational studies.

3.2. Regression Discontinuity

In the study of causal effects based on observational data, the Regression Discontinuity Design (RDD) offers certain advantages in addressing policy implementation issues with prescribed thresholds. This method was first formally introduced by Thistlethwaite and Campbell, and subsequent research has provided both theoretical support and empirical application [12]. Campbell further clarified the concept of RDD, laying the foundation for its application in policy evaluation [45]. The basic idea of RDD is to divide observations into experimental and control groups based on whether a specific variable reaches or exceeds a predetermined threshold. This threshold can be government-mandated, naturally occurring, or selected by researchers based on theoretical frameworks and data characteristics. By comparing observations near the threshold, researchers can more accurately estimate the impact of policy implementation on the dependent variable (Lee, 2008) [46].

RDD can be divided into two main types based on whether the treatment assignment is entirely determined by the running variable: Sharp Regression Discontinuity and Fuzzy Regression Discontinuity. The core feature of Sharp Regression Discontinuity is that the treatment assignment is solely determined by the running variable. When the running variable reaches a specific precise threshold, the assignment of the policy treatment changes abruptly, resulting in significant changes for individuals within the value range near the threshold. This design allows researchers to more accurately observe changes in causal effects at the threshold but also demands higher precision in data and threshold selection. Fuzzy Regression Discontinuity, on the other hand, allows for some degree of ambiguity in treatment assignment, meaning that the assignment of treatment does not entirely depend on the running variable when it reaches

a certain threshold. In this case, the assignment of treatment for individuals near the threshold may show a smoother transition rather than an abrupt change. The design of Fuzzy Regression Discontinuity is more aligned with the practical implementation of policies, allowing researchers to consider a certain degree of flexibility in treatment assignment. The choice between these two RDD designs in research depends on the researchers' theoretical assumptions about changes in causal effects and the availability and quality of data. Sharp Regression Discontinuity is suitable for studies with a high focus on specific changes in causal effects at the threshold, while Fuzzy Regression Discontinuity allows for some degree of flexibility in treatment assignment, better reflecting some uncertainties in actual policy implementation.

The application of RDD in causal inference reveals several important limitations. Firstly, with precise control over the treatment assignment indicator, the estimation of RDD may fail, thus compromising the reliability of the method to some extent. Secondly, this method can only identify "local" treatment effects near the threshold, imposing certain limitations on overall external validity. To overcome these limitations, Angrist and Rokkanen (2015) introduced the conditional independence assumption, drawing on the idea of matching methods to extend causal effects to any location on the threshold [47].

To further enhance the applicability of the RDD method, scholars have gradually expanded and improved the methodology. This process includes, but is not limited to, Multi-Running Variable RDD (Papay et al., 2011) [48], Quantile RDD (Frandsen et al., 2012) [49], Identification of Treatment Effects Away from the Threshold (Wing and Cook, 2013) [50], Multi-Threshold RDD (Cataneo et al., 2013) [51], Bending Regression Design (Card and Giuliano, 2016) [52], and Discrete Running Variable RDD (Kolesar and Rothe, 2018) [53]. The introduction of these extended methods has broadened the applicability of RDD in more complex practical scenarios, providing researchers with more flexible and diverse choices and tools. The continuous development of these methods offers a richer and deeper analytical framework for causal effect research, providing more comprehensive methodological support for solving practical research problems.

3.3. Propensity Score

In practical policy evaluation, propensity scores are introduced to address issues of selection bias and confounding. The core idea of propensity scores is to achieve covariate balance between treatment and control groups by developing a model to predict the probability of an individual being assigned to the treatment group. This concept essentially simulates a randomized experiment, thereby effectively eliminating confounding effects arising from individual characteristics. This method estimates the probability of each individual receiving the policy (i.e., the propensity score). For individual *i*, the propensity score is represented as $\pi(X_i;\alpha)$, where X_i denotes the individual's characteristics or prior factors. This score reflects the probability that the individual is exposed to the policy.

$$\pi(X_i;\alpha) = P(T_i = 1 | X_i;\alpha) \tag{2}$$

In this context, T_i is an indicator variable representing whether an individual has received policy intervention. By utilizing propensity scores, two primary methods can be employed to mitigate selection bias:

Matching: In the context of policy evaluation under non-randomized conditions, traditional Ordinary Least Squares (OLS) estimation methods struggle to address self-selection issues. Consequently, Propensity Score Matching (PSM) has become a widely adopted approach for

ISSN: 1813-4890

minimizing selection bias in non-experimental studies (Caliendo and Kopeinig, 2008) [54]. PSM adopts the concept of simulating random assignment by estimating the probability of each individual receiving the policy (i.e., the propensity score) and then matching individuals who received the policy with those who did not but have similar propensity scores. This matching process results in the creation of a control group akin to a randomized trial, thereby facilitating a better estimation of the causal effect of the policy. Rosenbaum and Rubin proposed that if the covariate vector satisfies the conditional independence assumption, then treating the propensity score as a function of the covariates can also ensure the establishment of conditional independence [55]. Thus, by matching based on the propensity score, unbiased estimates of treatment effects can be obtained at the individual level. The propensity score possesses a balancing property, such that individuals with the same propensity score across different groups tend to have similar covariate distributions.

$$P\left(X_i = x_i \mid PS_i, D_i = 1\right) = P\left(X_i = x_i \mid PS_i, D_i = 0\right) \tag{3}$$

This implies that the covariate distribution between the treated and untreated groups tends to approximate randomization, thereby simplifying the estimation of the treatment effect. PSM also facilitates the transformation of a multidimensional covariate matching problem into a one-dimensional propensity score matching by integrating multiple covariates into a single propensity score, thereby identifying similar control group individuals and generating counterfactuals for the treatment group (Bryson et al., 2002; Oh et al., 2009; Zhang Menglin and Li Guoping, 2020) [56-58].

However, the matching method has its limitations. It can only match based on observable variables and struggles to control for heterogeneity arising from unobservable factors. Moreover, the covariate distribution of the matched sample may not align with the overall population distribution, thus limiting the generalizability of the estimated causal effects to the entire population and providing estimates only for a smaller subpopulation. The implementation of matching relies on the non-confounding assumption. Sensitivity analyses of this assumption indicate that if unobservable confounders exist, propensity score matching may fail to eliminate selection bias and could potentially introduce new biases (Heckman, 2008; Liu Fengqin and Ma Hui, 2009; Damrongplasit, 2010) [59-61]. Therefore, it is crucial to carefully consider the limitations when applying the matching method.

Weighting: Propensity scores are not only tools for covariate balancing but are also regarded as a dimension reduction technique. Beyond balancing covariate distributions between groups, they can be applied to sample weighting adjustments. In this context, Inverse Probability Weighting (IPW) is a commonly used technique. IPW, akin to standardization, has its origins in survey research over sixty years ago and was later introduced to causal inference in observational studies by Robins et al. [62]. This method generates weights based on the inverse probability of receiving treatment, assigning weights to each individual to eliminate imbalances between treated and untreated individuals, thereby reducing selection bias in non-experimental studies and providing unbiased causal effect estimates (Rosenbaum, 1987; Stuart et al., 2014) [63, 64]. This approach emphasizes individuals with greater differences in prior factors, particularly when estimating overall effects (Linden and Adams, 2010) [65]. Propensity score methods offer a flexible and robust tool for policy evaluation, but their application requires careful consideration of model assumptions and the selection of appropriate control variables to ensure the validity of causal inferences (Athey and Imbens, 2017; Ritz, 2011; Coursey et al., 2012) [66, 67, 68]. In practice, a pretreatment model (typically a logistic regression model) is first used to fit the probability of each individual being assigned to the treatment group (i.e., the propensity score).

ISSN: 1813-4890

Based on the balancing property of the propensity score, the probability of an individual entering the treatment group or the control group is the same. This results in the propensity score satisfying the moment equation $E[Z-\pi(Z;\alpha X)]=0$. The unknown parameter $\alpha\alpha$ is estimated using Generalized Method of Moments (GMM). Once $\alpha\alpha$ is obtained, inverse propensity scores are used as weights. Specifically, for individuals who received the treatment, the weight is $1/\pi(X_i;\hat{\alpha})$, and for those who did not receive the treatment, the weight is $1/(1-\pi(X_i;\hat{\alpha}))$. The inverse probability weighted estimate of the average causal effect is given by:

$$\hat{\tau}_{IPW}=\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\cdot D_i}{\pi\left(X_i;\hat{\alpha}\right)}-\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\cdot\left(1-D_i\right)}{1-\pi\left(X_i;\hat{\alpha}\right)} \tag{4}$$

This method assumes that the treatment probability model is correct, and the weighted sample after inverse probability weighting can be considered random (Hernán and Robins, 2016) [69]. In practical applications, it is challenging to ascertain whether the treatment probability model encompasses all key covariates. The selection of covariates can influence the propensity score, which in turn affects the stability of the inverse probability weighted estimates. To mitigate this uncertainty, different approaches can be employed to construct the propensity score and inverse probability, aiming to enhance the stability of the estimates (Robins and Hernan, 2000; Hirano and Imbens, 2001) [70, 71].

3.4. Double Robust Estimation

The Double Robust Estimation (DRE) method has played a significant role in policy evaluation, particularly in addressing causal inference or correcting potential selection bias. The core objective of policy evaluation is to accurately estimate the causal effect of a policy or treatment on a specific outcome. However, due to the non-random nature of assignment, there may be underlying differences between the treatment and control groups, posing challenges for reliable estimation. In this context, the DRE method offers a more robust estimation strategy for causal inference in non-randomized settings by cleverly combining propensity score models and outcome models.

The central idea of DRE is to adjust for selection bias between the treatment and control groups through the propensity score model. First, methods such as logistic regression are used to estimate the propensity score $\hat{e}\left(X_i\right)$, i.e., the probability of being in the treatment group. Subsequently, techniques such as weighted regression estimation or bias-corrected weighting are employed to incorporate the propensity score into the outcome model, thereby obtaining a double robust estimate of the causal effect. This gives the estimator its double robustness property: as long as either the regression model or the propensity score model is correctly specified, the double robust estimator remains consistent. The estimation formula is as follows:

$$\hat{\tau}_{DR}=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i-\hat{e}\left(X_i\right)}{\hat{e}\left(X_i\right)\cdot\left(1-\hat{e}\left(X_i\right)\right)}\cdot\left(Y_i-\hat{m}\left(X_i,D_i\right)\right)+\hat{m}\left(X_i,1\right)-\hat{m}\left(X_i,0\right)\right] \tag{5}$$

In this context, $\hat{m}\left(X_i,D_i\right)$ represents the estimate of the outcome variable Y_i given the covariates X_i and the treatment variable D_i . The application of this method in policy evaluation allows researchers to more accurately assess policy effects while enhancing resistance to potential selection bias, thereby increasing the credibility of the estimation results. By comprehensively considering the differences between the treatment and control groups as well

ISSN: 1813-4890

as the fluctuations in the outcome variable, double robust estimation provides a powerful and flexible tool for causal inference, offering more reliable data support for policy-making.

Double robust estimation methods come in various forms, including weighted regression estimation and bias-corrected weighting. The core idea of these methods is to introduce weights into the modeling of both the propensity score and the outcome variable to more effectively address potential selection bias and other estimation biases.

Weighted Regression Estimation: In this approach, both the propensity score model and the outcome model can be estimated using weighted regression. Typically, the choice of weights is based on the inverse probability of the propensity score, to give more focus to observations that are less balanced in terms of treatment selection. When computing the double robust estimate, weights are used to adjust for the differences between the treatment and control groups, thereby improving the accuracy of the estimation (Robins and Rotnitzky, 2001; Tsiatis, 2006; Schafer and Kang, 2008) [72, 73, 74].

Bias-Corrected Weighting: This method usually involves using weights to correct for estimation bias. A common implementation is to use the inverse probability weights of the propensity score to adjust for the differences between the treatment and control groups, coupled with some correction terms to reduce the estimation bias introduced by the weights (Abadie and Imbens, 2011) [75]. The choice of method depends on the specific context of the study, the nature of the data, and the researchers' assumptions about potential biases.

Weighted regression estimation and bias-corrected weighting represent two main variants of double robust estimation, both aiming to enhance the robustness and efficiency of causal inference. In practical applications, researchers need to select the most appropriate method based on the context of the problem and the characteristics of the data to ensure that the estimation results are more credible and generalizable.

3.5. Differences-in-Differences Method

In the field of policy evaluation, the Differences-in-Differences (DID) method is an important econometric technique widely used to estimate the causal effect of a policy on specific groups or entities (Athey and Imbens, 2006; Dimick and Ryan, 2014) [76, 77]. The core idea of this method is to eliminate potential confounding factors that could bias the evaluation results by comparing the differences between the group subjected to the policy intervention and the group not subjected to the policy intervention before and after the policy implementation. The specific formula for the DID method in policy evaluation is as follows:

$$Y_{it} = \beta_0 + \beta_1 \times \mathrm{T}_i + \beta_2 \times \mathrm{P}_t + \beta_3 \times \left(\mathrm{T}_i \times \mathrm{P}_t\right) + \varepsilon_{it} \tag{6}$$

In the provided text, Y_{it} denotes the observed value for observation unit i at time t; T_i is a binary variable indicating whether the observation unit has been subjected to policy intervention. When a policy intervention takes place, T_i equals 1; otherwise, it is 0. Similarly, P_t is a binary variable indicating whether the observation time is post-policy implementation. At time points subsequent to the policy implementation, P_t is set to 1; otherwise, it is 0.

The crux of this formula lies in the interaction term $\left(\mathrm{T}_i \times \mathrm{P}_t\right)$,which encapsulates the differential changes between the group impacted by the policy post-implementation and the unaffected group. If β_3 is significantly different from zero, it suggests that the policy has elicited a significant causal effect. The intercept β_0 represents the average observed value of the group unaffected by the policy prior to its implementation. The coefficient β_1 signifies the difference

in intercepts between the group affected by the policy before implementation and the unaffected group. The coefficient β_2 reflects the average policy impact across all groups post-implementation. The interaction term β_3 indicates the differential policy effects between the group affected by the policy post-implementation and the unaffected group. The DID method, by comparing changes pre- and post-policy implementation between the two groups, controls for time trends and individual heterogeneity, thereby enhancing the accuracy of policy effect estimation. The strength of this method lies in its relatively straightforward yet robust framework for estimating causal effects, making it suitable for numerous policy evaluation issues in empirical research.

In practical applications, researchers ascertain the estimated effect of the DID method by comparing differences between groups before and after policy implementation (Athey and Imbens, 2017) [66]. The DID method has emerged as a potent tool in policy evaluation, particularly in scenarios where constructing a control group is challenging or where endogeneity issues arise. However, to ensure the reliability of results, researchers must validate the core assumptions of the DID method, particularly the parallel trends assumption. The application of the DID method in policy evaluation offers a robust and flexible analytical tool for a profound understanding of policy impacts. (Fan Dan et al., 2017) utilized a kernel matching difference-in-differences model, controlling for variables such as regional environmental regulation, to assess the impact of China's pilot carbon emission trading rights on low-carbon economic transformation. The findings indicate that the carbon emission trading mechanism is highly effective in fostering technological innovation, providing essential measures for achieving a low-carbon economic transformation [78]. This study applied DID and policy evaluation methods to deeply analyze the policy effects of the carbon emission trading pilot. Yin Zhichao and Guo Peiyao employed the DID method to evaluate the impact of the targeted poverty alleviation policy on the consumption of impoverished families. The research revealed that the policy significantly increased per capita consumption levels, particularly in subsistence and developmental consumption. The results underscore the policy's significant positive effect in deeply impoverished areas, offering a robust reference for future poverty alleviation policies [79].

3.6. Synthetic Control Method and Regression Synthetic Method

The Synthetic Control Method (SCM) and Regression Synthetic Method (RSM) play crucial roles in the field of policy evaluation, particularly in addressing potential selection biases in observational data. The SCM constructs a "synthetic" control group by taking a weighted average of observed control group units to create an ideal control group for the treatment group. This method requires that, prior to the intervention, the outcome and covariate variables of the synthetic control group closely resemble those of the treated units, ensuring more robust estimates of the treatment effect. In policy evaluation, SCM helps researchers simulate an ideal control group within observational data, thereby allowing for a more accurate assessment of the true effect of the policy.

In contrast, the RSM imposes less stringent constraints on the weights, allowing for negative weights. This method estimates the post-treatment counterfactual outcomes for treated units by considering the correlations among individuals, particularly common factors. The flexibility of RSM in policy evaluation enables the use of more complex models to address challenges in causal inference, especially when there are significant differences between the treatment and control groups, making it better suited to complex data structures.

To address the robustness issue of SCM when there is only one treated unit, Xu introduced panel interactive fixed effects into the causal inference framework, proposing the Generalized Synthetic Control Method. This method extends Abadie's SCM to multiple treatment groups and multi-period policies [80]. In this framework, SCM and linear fixed effects models are connected,

with the DID method being a special case. Bai and colleagues extended the RSM to non-stationary data contexts, where the Ordinary Least Squares (OLS) estimates of the related coefficients are consistent [81]. Doudchenko and Imbens explored alternative methods for computing appropriate weights for SCM, such as Best Subset Regression, LASSO, and Elastic Nets, which perform better when there are a large number of control units [82].

These extensions have enriched the application scope of SCM and RSM, making them more suitable for complex and diverse policy evaluation scenarios. Researchers can more flexibly choose and apply these methods when dealing with various treatment groups and multi-period policies, thereby better addressing the challenges of non-experimental data, and improving the accuracy and robustness of causal inference.

4. The Research Value and Future Trends of Policy Evaluation

In contemporary times, social research is undergoing a paradigm shift from statistical inference to causal inference (Panhas and Singleton, 2017) [83]. This signifies an increasing methodological emphasis among researchers on pursuing causal relationships rather than merely relying on statistical inference. This shift provides a new research perspective that deepens the understanding of social phenomena. Policy evaluation based on causal inference has emerged as a pivotal method in both research and practice. Its core lies in employing scientific research design and data analysis to more accurately measure and assess the actual impact of policies on specific outcomes.

4.1. The Research Value of Policy Evaluation

Causal inference plays a pivotal role in policy evaluation, not only driving advancements in existing causal inference theories at the theoretical level but also significantly promoting practical applications.

From a theoretical perspective, causal inference provides a clear and systematic analytical framework for policy evaluation, laying a solid foundation for our deep understanding of the impact of policy implementation on social phenomena (Angrist and Pischke, 2008)[84]. By constructing models of causal relationships, we can delve into the association between policy measures and specific outcomes, effectively avoiding potential confounding and misinterpretations in causal relationships (Imbens and Rubin, 2015)[85]. This theoretical value underscores the importance of considering endogeneity issues in research to ensure the robustness of scientific, valid, and reliable research conclusions. Causal inference offers a concrete methodology for policy evaluation, providing researchers with powerful tools to analyze policy effects in complex social environments.

In practice, causal inference equips policymakers with practical tools, offering decision-makers more specific and actionable data support. Employing empirical methods such as randomized controlled trials (Schultz and Strauss, 2008; Deaton, 2009)[86, 87], we can more precisely quantify the actual effects of policies, providing clear guidance for policy adjustment and optimization. This empirical analysis not only aids in quantifying the magnitude of policy effects but also delves into potential influencing factors, offering comprehensive policy recommendations. On the other hand, causal inference provides decision-makers with scientific and actionable information, enhancing the effectiveness of policy implementation, making policies more aligned with societal needs, and achieving long-term sustainability goals. Thus, the integration of theory and practice enables causal inference to play a dual and synergistic role in the field of policy evaluation, providing a solid foundation for more effective decision-making.

ISSN: 1813-4890

4.2. Future Trends in Policy Evaluation

Although numerous scholars have focused on and utilized causal inference for policy evaluation research, it is evident that several unresolved issues persist in current policy evaluations. Through the literature review presented earlier, we can observe that past research has primarily concentrated on binary treatment of single policies, evaluating the implementation or non-implementation of a policy to reveal its impact on specific outcomes. However, in practice, the implementation of policies often involves the intersection of multiple policies, rendering the binary treatment of a single policy insufficient to comprehensively reflect the complex effects of policy interactions.

As researchers confront the challenges posed by the interplay of multiple policies, it has become apparent that more comprehensive policy evaluation methods are necessary. This includes considering the joint effects and interaction effects that may arise when multiple policies coexist, to more accurately understand the combined impact of these policies on outcomes. To address this challenge, the evaluation of causal effects with multivariable treatments is emerging as a new direction in research. The goal of this approach is to gain a more nuanced understanding of the complexity of policy implementation and the multi-layered interaction effects, thus providing a more detailed assessment of policy impacts.

While policy evaluation remains in a stage of ongoing exploration, many scholars engaged in this field are actively investigating the theoretical value and practical applications of causal inference in policy evaluation. This paper will focus on discussing three key aspects to explore potential future trends in policy evaluation research under the framework of causal inference, aiming to offer insights for deeper investigations in this field.

4.2.1. From Panel Data to Time Series Data

From the perspective of data structure, the study of causal inference is undergoing a significant evolution, transitioning from cross-sectional and short panel data (Wooldridge, 2002)[88] to time series data (Angrist and Pischke, 2009)[89]. This shift underscores an enhanced focus on the temporal dimension, reflecting researchers' desire for a more comprehensive understanding of causal relationships and policy effects. As this trend intensifies, time series data has increasingly become the focal point in the field of causal inference, as it provides richer and more dynamic information, enabling researchers to more accurately capture the relationships between variables, particularly the changes in causal relationships over time (Hansen, 2008)[90].

In the realm of policy evaluation, the application of time series data is becoming crucial. Policies often do not yield immediate effects but gradually manifest their impacts over time. By employing time series data, researchers can track the process of policy implementation, capturing the gradual emergence and evolution of policy effects, thereby gaining a more comprehensive understanding of the long-term impacts and trends of policies. The use of time series data allows researchers to better comprehend the dynamic relationships between variables, revealing the causal pathways that evolve over time with policy changes. This dynamic perspective provides a deeper understanding for policy evaluation, aiding in more accurately quantifying the long-term impacts of policies.

4.2.2. Transformation in Policy Evaluation Methods

In the ongoing evolution of causal inference methodologies, researchers are gradually shifting their focus from the traditional emphasis on average treatment effects to an in-depth exploration of heterogeneous treatment effects (Angrist and Pischke, 2009)[89]. The average treatment effect, as an aggregate measure, may sometimes obscure the heterogeneous effects that exist among different individuals, groups, or environments. Therefore, a greater emphasis on heterogeneous treatment effects is crucial for accurately understanding the impact of policies on different groups or under specific conditions, thereby enhancing the

comprehensiveness and granularity of policy evaluations. Delving into heterogeneous treatment effects helps to uncover variations in policy outcomes across different contextual conditions, providing deeper insights for policymakers to tailor policies more effectively (Abadie et al., 2010)[91]. This shift towards focusing on heterogeneity enables researchers to dissect the diversity in policy impacts more precisely, offering more specific and effective recommendations for practical policy formulation. By thoroughly investigating heterogeneous treatment effects, researchers can better understand the varying responses of different groups during policy implementation, thereby providing a more precise and personalized analytical framework for policy evaluation.

On the other hand, research methodologies are gradually transitioning from the analysis of static treatment effects to more in-depth analyses of dynamic treatment effects. This transformation reflects a deeper recognition of the complexities involved in policy impacts, as the actual effects of policies often evolve over time. In the context of policy evaluation, an increased emphasis on dynamic treatment effects analysis aids in comprehensively understanding the impact pathways and trends of policies over different time periods (Fredriksson and Johansson, 2008; Abadie et al., 2010)[92, 91]. As time progresses, policy effects may exhibit diverse patterns and dynamic characteristics. Static treatment effects sometimes fail to capture these changes; thus, focusing on dynamic treatment effects allows researchers to more accurately describe the actual impacts of policies at different temporal stages. This dynamic analysis helps to reveal the temporal evolution of policy effects, providing deeper insights and making policy evaluations more practically informative.

Moreover, the study of dynamic treatment effects not only captures the temporal variations in policy impacts but also helps to understand potential lagged effects, cumulative effects, and the interactions of policies at different stages. This detailed temporal analysis enables researchers to gain a more comprehensive understanding of the long-term impacts of policies, offering more accurate and practical information for decision-makers. This methodological transformation provides more comprehensive and in-depth tools for policy evaluation, promising to advance causal inference research towards a more practice-oriented and policy-making-focused direction.

4.2.3. Integration of Machine Learning and Big Data in Policy Evaluation under Causal Inference

The integration of machine learning and big data in policy evaluation under causal inference is flourishing. With the advent of the big data era, researchers in policy evaluation increasingly recognize the immense potential of machine learning in handling large-scale, high-dimensional data (Chernozhukov et al., 2018)[93]. This development is driven by the flexible adaptability and robust performance of machine learning algorithms in addressing challenging issues in policy effect evaluations.

Firstly, big data provides a wealth of extensive information, enabling policy evaluation to transcend the limitations of small samples or specific domain data. Machine learning algorithms can process these large datasets, effectively extracting patterns, relationships, and trends, laying a more comprehensive foundation for the accurate assessment of policy effects. Moreover, the complex relationships in high-dimensional data often exceed the capabilities of traditional statistical methods, while machine learning's nonlinear, nonparametric approaches are better equipped to capture this complexity.

Secondly, machine learning is increasingly prominent in causal inference. Traditional causal inference methods may struggle with multicollinearity, interaction effects, and nonlinear relationships, whereas machine learning algorithms can more flexibly adapt to these scenarios. For example, methods like causal forests and inference neural networks offer new avenues for handling complex causal relationships in the context of big data, aiding in more accurate

estimation of policy effects (Wager and Athey, 2018)[94]. However, the development in this field also faces challenges. The black-box nature of machine learning and the lack of interpretability may make the interpretation of causal relationships more difficult. Therefore, researchers need to strike a balance between strong predictive performance and interpretability to ensure the application of machine learning in policy evaluation is credible and interpretable.

As China continues to advance its reform and opening-up, the government is implementing a series of proactive policy measures. In this context, policy evaluation under causal inference is entering a new phase of development. The evolution of policy evaluation research is gradually moving beyond merely introducing causal inference methods to assess policy implementation effects, and is increasingly focusing on how to skillfully apply causal inference techniques to more deeply guide the future development direction of our country.

This new phase will emphasize a more comprehensive and profound understanding of policy effects, transcending simple effect quantification. By leveraging causal inference techniques, policy evaluation research will focus more on revealing the impacts of policy implementation on different groups, at different time stages, and across different regions, providing decision-makers with more accurate and specific policy recommendations. This means that policy evaluation will pay more attention to solving practical problems, driving more significant achievements in various fields in our country. In the future, the use of causal inference techniques will not merely be a means to evaluate policies but will become an intelligent tool guiding our country's development direction. Policy makers and researchers will work together to deeply explore the potential of causal inference techniques, better understanding the long-term impacts of policy changes, and providing more scientific and reliable support for national strategies and decisions. This new stage will drive the deep integration of policy evaluation with causal inference technology, making it an important intellectual support system shaping our country's future.

References

[1] Ronald Aylmer Fisher et al. The Design of Experiments, (7th Ed), 1960.

[2] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. Randomized Controlled Trials. AJR Am JRoentgenol, 183(6):1539–44, 2004.

[3] Herman Wold. Causality and econometrics. Econometrica: Journal of the Econometric Society, pages162–177, 1954.

[4] Kevin D Hoover. The logic of causal inference: Econometrics and the conditional analysis of causation.Economics & Philosophy, 6(2):207–234, 1990.

[5] David Roxbee Cox. Planning of experiments. 1958.

[6] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies.Journal of educational Psychology, 66(5):688, 1974.

[7] James J Heckman. Sample selection bias as a specification error. Econometrica: Journal of the econometric society, pages 153–161, 1979.

[8] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. Journal of the American statistical Association, 91(434):444–455, 1996.

[9] Judea Pearl et al. Models, reasoning and inference. Cambridge, UK: Cambridge University Press, 19(2):3, 2000.

[10] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causalstudies. Review of Economics and statistics, 84(1):151–161, 2002.

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

[12] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternativeto the ex post facto experiment. Journal of Educational psychology, 51(6):309, 1960.

[13] Jeff Wooldridge and Guido Imbens. Difference-in-differences estimation. Lecture notes, 10, 2007.

[14] Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensityscore for estimating the average treatment effect of multiple treatments. Statistics in medicine,31(7):681–697, 2012.

[15] Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. Journal of econometrics, 142(2):615–635, 2008.

[16] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. American journal of epidemiology, 173 (7):761–767, 2011.

[17] Daron Acemoglu and Joshua Angrist. How large are human-capital externalities? evidence from compulsory schooling laws. NBER macroeconomics annual, 15:9–59, 2000.

[18] Joshua Angrist and William N Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size, 1996.

[19] Eline Sneyers and Kristof De Witte. Interventions in higher education and their effect on student success: A meta-analysis. Educational Review, 70(2):208–228, 2018.

[20] Rob Greenwald, Larry V Hedges, and Richard D Laine. The effect of school resources on student achievement. Review of educational research, 66(3):361–396, 1996.

[21] Melissa Dell. The persistent effects of peru's mining mita. Econometrica, 78(6):1863–1903, 2010.

[22] Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. Proceedings of the National Academy of Sciences, 110(32):12936–12941, 2013.

[23] Vincent Vandenberghe and Stephane Robin. Evaluating the effectiveness of private education across countries: a comparison of methods. Labour economics, 11(4):487–506, 2004.

[24] Qi Zhanyong; Du Yue. Assessment of the Impact of Educational Policy Implementation. Educational Research.44(145).

[25] Shu Yang, Guido W Imbens, Zhanglin Cui, Douglas E Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics,72(4):1055–1065, 2016.

[26] Sylvain Chabé-Ferret and Julie Subervie. How much green for the buck? estimating additional and windfall effects of french agro-environmental schemes by did-matching. Journal of Environmental Economics and Management, 65(1):12–27, 2013.

[27] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. Ann. Agricultural Sciences, pages 1–51, 1923.

[28] Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics, pages 185–203, 1973.

[29] Donald B Rubin. Assignment to treatment group on the basis of a covariate. Journal of educational Statistics, 2(1):1–26, 1977.

[30] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. The Annals of statistics, pages 34–58, 1978.

[31] Sophie Webber and Carolyn Prouse. The new gold standard: The rise of randomized control trials and experimental development. Economic Geography, 94(2):166–187, 2018.

[32] Keith Morrison. Randomised controlled trials for evidence-based education: some problems in judging'what works'. Evaluation & Research in Education, 15(2):69–83, 2001.

[33] Carole J Torgerson and David J Torgerson. The need for randomised controlled trials in educational research. In Education matters, pages 203–214. Routledge, 2012.

[34] Paul Connolly, Ciara Keenan, and Karolina Urbanska. The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016.Educational Research, 60(3):276–291, 2018.

[35] Maureen Dobbins, Steven E Hanna, Donna Ciliska, Steve Manske, Roy Cameron, Shawna L Mercer,Linda O'Mara, Kara DeCorby, and Paula Robeson. A randomized controlled trial evaluating the impact of knowledge translation and exchange strategies. Implementation science, 4(1):1–16, 2009.

[36] Simon Mark Murphy, Rhiannon Tudor Edwards, Nefyn Williams, Larry Raisanen, Graham Moore,Pat Linck, Natalia Hounsome, Nafees Ud Din, and Laurence Moore. An evaluation of the effectiveness and cost effectiveness of the national exercise referral scheme in wales, uk: a randomised controlled trial of a public health policy initiative. J Epidemiol Community Health, 2012.

[37] Chris Bonell, Adam Fletcher, Matthew Morton, Theo Lorenc, and Laurence Moore. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. Social science & medicine, 75(12):2299–2306, 2012.

[38] Gary R Bond, Robert E Drake, and Deborah R Becker. An update on randomized controlled trials of evidence-based supported employment. Psychiatric rehabilitation journal, 31(4):280, 2008.

[39] María Laura Alzúa, Guillermo Cruces, and Carolina Lopez Erazo. Youth training programs beyond employment. evidence from a randomized controlled trial. La Plata, Buenos Aires: CEDL AS. (Mimeo.), 2013.

[40] Joanne Wall, Cliona Ni Mhurchu, Tony Blakely, Anthony Rodgers, and Jenny Wilton. Effectiveness of monetary incentives in modifying dietary behavior: a review of randomized, controlled trials.Nutrition reviews, 64(12):518–531, 2006.

[41] Cliona Ni Mhurchu, Tony Blakely, Yannan Jiang, Helen C Eyles, and Anthony Rodgers. Effects of price discounts and tailored nutrition education on supermarket purchases: a randomized controlled trial. The American journal of clinical nutrition, 91(3):736–747, 2010.

[42] Gary S Bilotta, Alice M Milner, and Ian Boyd. On the use of systematic reviews to inform environmental policies. Environmental Science & Policy, 42:67–77, 2014.

[43] Elisha R Frederiks, Karen Stenner, Elizabeth V Hobman, and Mark Fischle. Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers. Energy research & social science, 22:147–164, 2016.

[44] Thomas Clasen, William Checkley, Jennifer L Peel, Kalpana Balakrishnan, John P McCracken, Ghislaine Rosa, Lisa M Thompson, Dana Boyd Barr, Maggie L Clark, Michael A Johnson, et al. Design and rationale of the hapin study: a multicountry randomized controlled trial to assess the effect ofliquefied petroleum gas stove and continuous fuel distribution. Environmental health perspectives,128(4):047008, 2020.

[45] Donald T Campbell and HW Riecken. Quasi-experimental design. International encyclopedia of the social sciences, 5(3):259–263, 1968.

[46] David S Lee. Randomized experiments from non-random selection in us house elections. Journal of Econometrics, 142(2):675–697, 2008.

[47] Joshua D Angrist and Miikka Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. Journal of the American Statistical Association, 110 (512):1331–1344, 2015.

[48] John P Papay, John B Willett, and Richard J Murnane. Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics, 161(2):203–207, 2011.

[49] Brigham R Frandsen, Markus Frölich, and Blaise Melly. Quantile treatment effects in the regression discontinuity design. Journal of Econometrics, 168(2):382–395, 2012.

[50] Coady Wing and Thomas D Cook. Strengthening the regression discontinuity design using additional design elements: A within-study comparison. Journal of Policy Analysis and Management, 32(4):853–877, 2013.

[51] Matias D Cattaneo, David M Drukker, and Ashley D Holland. Estimation of multivalued treatment effects under conditional independence. The Stata Journal, 13(3):407–450, 2013.

[52] David Card and Laura Giuliano. Can tracking raise the test scores of high-ability minority students? American Economic Review, 106(10):2783–2816, 2016.

[53] Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8):2277–2304, 2018.

[54] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1):31–72, 2008.

[55] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

[56] Alex Bryson, Richard Dorsett, and Susan Purdon. The use of propensity score matching in the evaluation of active labour market policies. 2002.

[57] Inha Oh, Jeong-Dong Lee, Almas Heshmati, and Gyoung-Gyu Choi. Evaluation of credit guarantee policy using propensity score matching. Small Business Economics, 33:335–351, 2009.

[58] Zhang Menglin and Li Guoping. Policy Effect Evaluation and Mechanism Analysis of Commercial Insurance in Reducing Household Poverty Vulnerability. Contemporary Economic Research. 303(11):91–102, 2020.

[59] James J Heckman. Econometric causality. International statistical review, 76(1):1–27, 2008.

[60] Liu Fengqin and Ma Hui. Sensitivity Analysis of Propensity Score Matching Method. Statistics and Information Forum 24(10):7–13, 2009.

[61] Kannika Damrongplasit, Cheng Hsiao, and Xueyan Zhao. Decriminalization and marijuana smoking prevalence: Evidence from australia. Journal of Business & Economic Statistics, 28(3):344–356, 2010.

[62] James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics, pages 479–495, 1992

[63] Paul R Rosenbaum. The role of a second control group in an observational study. Statistical Science,2(3):292–306, 1987.

[64] Elizabeth A Stuart, Haiden A Huskamp, Kenneth Duckworth, Jeffrey Simmons, Zirui Song, Michael E Chernew, and Colleen L Barry. Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology, 14:166–182, 2014.

[65] Ariel Linden and John L Adams. Using propensity score-based weighting in the evaluation of healthmanagement programme effectiveness. Journal of Evaluation in Clinical Practice, 16(1):175–179,2010.

[66] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. Journal of Economic perspectives, 31(2):3–32, 2017.

[67] Adrian Ritz. Attraction to public policy-making: A qualitative inquiry into improvements in psm measurement. Public Administration, 89(3):1128–1147, 2011.

[68] David Coursey, Kaifeng Yang, and Sanjay K Pandey. Public service motivation (psm) and support for citizen participation: A test of perry and vandenabeele's reformulation of psm theory. Public Administration Review, 72(4):572–582, 2012.

[69] Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. American journal of epidemiology, 183(8):758–764, 2016.

[70] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. Epidemiology, pages 550–560, 2000.

[71] Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes research methodology, 2:259–278, 2001.

[72] James M Robins and Andrea Rotnitzky. Comment on the bickel and kwon article,"inference for semiparametric models: Some questions and an answer". Statistica Sinica, 11(4):920–936, 2001.

[73] Anastasios A Tsiatis. Semiparametric theory and missing data. 2006.

[74] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychological methods, 13(4):279, 2008.

[75] Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics, 29(1):1–11, 2011.

[76] Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. Econometrica, 74(2):431–497, 2006.

[77] Justin B Dimick and Andrew M Ryan. Methods for evaluating changes in health care policy: the difference-in-differences approach. Jama, 312(22):2401–2402, 2014.

[78] Fan Dan, Wang Weiguo, and Liang Peifeng. Policy Effect Analysis of China's Carbon Emission Trading Mechanism: An Estimation Based on the Difference-in-Differences Model. Chinese Journal of Environmental Science.37(6):2383–2392, 2017.

[79] Yin Zhichao and Guo Peiyao. Evaluation of the Poverty Alleviation Policy Effects — An Empirical Study from the Perspective of Household Consumption. Management World.4:64–83,2021.

[80] Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis, 25(1):57–76, 2017.

[81] ChongEn Bai, Qi Li, and Min Ouyang. Property taxes and home prices: A tale of two cities. Journal of Econometrics, 180(1):1–15, 2014.

[82] Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

[83] Matthew T Panhans and John D Singleton. The empirical economist's toolkit: from models to methods. History of Political Economy, 49(Supplement):127–157, 2017.

[84] Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion princeton university press; 2008.

[85] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences.Cambridge University Press, 2015.

[86] T Paul Schultz and John Strauss. Handbook of development economics, volume 4. Elsevier, 2008.

[87] Angus S Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report, National bureau of economic research, 2009.

[88] Jeffrey M Wooldridge. Econometric analysis of cross section and panel data mit press. Cambridge, ma, 108(2):245–254, 2002.

[89] Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton university press, 2009.

[90] Ben B Hansen. The prognostic analogue of the propensity score. Biometrika, 95(2):481–488, 2008.

[91] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. Journal of the American statistical Association, 105(490):493–505, 2010.

[92] Peter Fredriksson and Per Johansson. Dynamic treatment assignment: the consequences for evaluations using observational data. Journal of Business & Economic Statistics, 26(4):435–445, 2008.

[93] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[94] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.