Efficient and Robust Estimation of the Generalized LATE Model

Haitian Xie
Guanghua School of Management, Peking University, Beijing, China

ABSTRACT
This article studies the estimation of causal parameters in the generalized local average treatment effect (GLATE) model, which expands upon the traditional LATE model to include multivalued treatments. We derive the efficient influence function (EIF) and the semiparametric efficiency bound (SPEB) for two types of causal parameters: the local average structural function (LASF) and the local average structural function for the treated (LASFT). The moment conditions generated by the EIF satisfy two robustness properties: double robustness and Neyman orthogonality. Based on the robust moment conditions, we propose the double/debiased machine learning (DML) estimator for estimating the LASF. The DML estimator is well-suited for high dimensional settings. We also propose null-restricted inference methods that are robust against weak identification issues. As an empirical application of these methods, we examine the potential health outcome across different types of health insurance plans using data from the Oregon Health Insurance Experiment.

1. Introduction
Since the pioneering work of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), the local average treatment effect (LATE) model has become a frequently employed tool for conducting causal inference studies. In contrast to the classical instrumental variable (IV) regression model, which assumes homogeneity of treatment effects, the LATE framework allows treatment effects to vary among individuals. Under the well-known monotonicity condition, the average treatment effect can be identified for a subset of individuals whose treatment choices align with changes in instrument levels.

While the current LATE model has gained significant popularity, it inherently carries a limitation in only supporting binary treatment variables. This limitation can be particularly restrictive in socioeconomic contexts characterized by prevalent endogeneity issues and multiple treatments. Examples of these scenarios include parents’ selection of diverse childcare options (Galindo 2021) and preschool programs (Kline and Walters 2016) for their children, high school students’ choices among types of higher education institutions (Mountjoy 2022), families’ relocation decisions to distinct neighborhoods in housing experiments (Aliprantis and Richter 2020; Pinto 2021), and individuals’ selection from a spectrum of health insurance alternatives (Finkelstein et al. 2012). In these instances, treatment levels are unordered and nonnumerical, thereby precluding analysis by conventional extensions of the LATE model (Angrist and Imbens 1995), which require that treatment choices are naturally ordered (such as years of schooling) and a meaningful aggregator can be formulated.

To apply the LATE model to the above contexts, researchers frequently need to redefine the treatment so that it comprises only two levels. Considering the original treatment levels as substitutes, the combined net effect of the merged binary treatment arises from a mixture of effects across various complier margins. Merging treatment levels in this way can result in the loss of heterogeneity in treatment effects across different levels, complicating program evaluation, and undermining the causal interpretation of the estimates.

This article addresses the above issues by extending the LATE framework to directly accommodate the multiplicity of treatment levels. We introduce the new framework as the generalized LATE (GLATE) model. In the GLATE model, the causal parameters we examine include the local average structural function (LASF) and local average structural function for the treated (LASFT), which represent the mean potential outcomes for specific subpopulations defined by their treatment choice behaviors. The identification of these causal parameters builds upon the binary LATE model’s results and is established based on the unordered monotonicity assumption introduced by Heckman and Pinto (2018).¹

The main contribution of this article is the derivation of efficient and robust econometric procedures for the GLATE model. First, we establish the identification results in a setting incorporating covariates. Second, we derive the efficient influence function (EIF) and the semiparametric efficiency bound

¹To distinguish it from the GLATE model, we refer to the LATE model studied by Imbens and Angrist (1994) as the “binary LATE model.”
(SPEB) for the identified parameters, subsequently verifying that the conditional expectation projection (CEP) estimator (e.g., Chen, Hong, and Tarozzi 2008) achieves the SPEB when constructed directly from the identification result. Third, we show that the moment conditions formulated from the EIF satisfy double robustness and Neyman orthogonality. These robust moment conditions are leveraged to construct double/debiased machine learning (DML) estimators (Chernozhukov et al. 2018), which are well-suited for high-dimensional settings. Fourth, we propose null-restricted test statistics for inference, demonstrating their robustness against weak identification issues.

Consider our empirical study for concreteness: we investigate the impact of various types of health insurance on health outcomes, using data from the Oregon Health Experiment (Finkelstein et al. 2012). The treatment variable is the type of insurance plan—Medicaid, non-Medicaid, or no insurance. The lottery that provides the opportunity to enroll in a Medicaid program serves as the instrument. We categorize individuals into different subpopulations based on their insurance choice behavior and explore their potential health outcome, that is, LASF. For instance, we analyze the subpopulation of individuals who switch from having no insurance to enrolling in Medicaid after receiving the enrollment option, as well as the subpopulation of individuals who switch from non-Medicaid insurance to Medicaid after receiving the enrollment option. We estimate the mean potential health outcomes for these subpopulations if they were to participate in Medicaid.

The approach of merging Medicaid and non-Medicaid into a single treatment level denoting insurance coverage, as executed by Finkelstein et al. (2012) using a binary LATE model, does not allow for the same depth of analysis as our method. By considering alternative insurance plans as distinct treatment levels, the GLATE model allows for a more comprehensive understanding of the causal links. In particular, it enables the identification of the average potential outcomes for more specific subpopulations, and facilitates the investigation of substitution effects among different insurance choices.

The rest of our article is organized as follows. The remaining part of this section discusses the literature. Section 2 introduces the GLATE model and the nonparametric identification results. Section 3 calculates the EIF and SPEB for the identified causal parameters. Section 4 discusses the robust moment conditions and introduces the DML procedure. Section 5 proposes robust inference procedures under weak identification. Section 6 presents the empirical study. The proofs for theoretical results in the main text are collected in Appendix A in the supplementary material. Appendix B contains simulation results. Appendix C provides additional results for parameters implicitly defined by overidentifying moment conditions.

1.1. Related Literature

The GLATE model provides a method for conducting causal inference in the presence of endogeneity for multivalued and unordered treatments. Initially established in Heckman and Pinto (2018), the identification of LASF (conditional on covariates) relies on the unordered monotonicity condition. Within the GLATE model, this unordered monotonicity condition serves as the natural counterpart to the original monotonicity condition introduced by Imbens and Angrist (1994). Other variations of monotonicity have been explored for discrete and ordered treatments (Angrist and Imbens 1995), as well as continuous treatments (Angrist, Graddy, and Imbens 2000). Recent studies have proposed weaker monotonicity conditions in the binary treatment model with multiple instruments, including vector monotonicity (Goff 2020) and partial monotonicity (Mogstad, Torgovitsky, and Walters 2021, 2023), though these approaches have yet to be extended to the multivalued treatment case. Lee and Salanié (2018) relaxed monotonicity in a setting with multivalued treatment and continuous instruments, thereby generalizing the results from the local IV approach of Heckman and Vytlacil (2005). More recently, Vazquez-Bare (2022) proposes a generalization of monotonicity in relation to spillover effects.

The literature on semiparametric efficiency for treatment effects begins with cases where unconfoundedness is assumed. In the context of binary treatments, Hahn (1998) and Hirano, Imbens, and Ridder (2003) determined the efficiency bounds for both the average treatment effect (ATE) and the average treatment effect on the treated (ATT). For situations involving multivalued treatments, studies such as Cattaneo (2010), Farrell (2015), Lee (2018), and Ao, Calonico, and Lee (2021) have calculated efficiency bounds and proposed efficient estimators for various causal parameters. In scenarios where unconfoundedness does not hold and instruments are available, Frölich (2007) derives the efficiency bound for the LATE parameter, while Hong and Nekipelov (2010) extend this approach to estimating parameters indirectly defined by moment restrictions. Within a broader framework that includes missing data, Chen, Hong, and Tarozzi (2008) examine semiparametric efficiency bounds and the efficient estimation of parameters identified through overidentifying moment restrictions. Nonetheless, there is currently a lack of theoretical studies on semiparametric efficient estimation in models addressing both endogeneity and unordered multiple treatment levels.

Various methods exist for determining the influence function of semiparametric estimators, as detailed in Chapter 3 of Bickel et al. (1993), Newey (1990, 1994), and Ichimura and Newey (2022). A general framework for constructing robust moment conditions using influence functions is provided by Chernozhukov et al. (2022). Using these robust moment conditions, Chernozhukov et al. (2018) develop DML estimators appropriate for high-dimensional scenarios, where Donsker properties are no longer necessary for the nuisance parameter space when establishing the asymptotics of the semi-parametric estimator.

For doubly robust estimation, an early contribution by Tan (2006) examines the binary LATE model. Okui et al. (2012) investigate doubly robust estimation of a parametric model that indexes the dependence of the outcome on the endogenous

---

2 Recently, Newey and Stouli (2021) give necessary and sufficient characterization for identification of average structural function of multiple treatment levels, where the usual unconfoundedness (full independence) is relaxed to mean independence.

3 If the treatment variable were continuous instead of discrete, the analysis would differ. For instance, Flores et al. (2012) and Huber et al. (2020) provide such methodological and empirical analyses.
treatment variable. Uysal (2015) presents doubly robust estimators within a multivalued treatment effect framework, assuming unconfoundedness. Singh and Sun (2022) explore double robustness of implicitly defined parameters for compliers in the binary LATE model. Słoczyński and Wooldridge (2018) offer a comprehensive overview of double robustness literature and a general identification result of double robustness.

For testing the GLATE model assumptions, Sun (2023) proposes a bootstrap test which is the generalization and improvement of the test suggested by Kitagawa (2015) in the binary LATE model.

2. Identification in the GLATE Model

In this section, we use the empirical study to motivate the introduction of the generalized local average treatment effect (GLATE) model. We then explore the identification of the local average structural function (LASF) and other related causal parameters.

2.1. The Model and Matrix Notation

The model comprises the following variables: instrument, treatment, outcome, and covariates. The instrument Z takes values in a finite set \( Z = \{z_1, \ldots, z_{N_Z}\} \). There is a collection of \( N_Z \) potential treatment variables, \( (T_1, \ldots, T_{N_Z}) \), each taking values in a finite set \( T = \{t_1, \ldots, t_{N_T}\} \). Each potential treatment \( T_z \) represents the treatment response that would have been observed if \( Z = z \). There is a collection of \( N_Y \) potential outcome variables, \( (Y_1, \ldots, Y_{N_Y}) \), each taking values in \( Y \subset \mathbb{R} \). Each \( Y_t \) represents the potential outcome that would have been observed if the treatment status were set at \( T = t \). The potential treatments and outcomes are not directly observed. Instead, we only observe the realized treatment and outcome, defined by \( T = T_Z = \sum_{z \in Z} I[Z = z]T_z \) and \( Y = Y_T = \sum_{t \in T} I[T = t]Y_t \), respectively. Additionally, there is a vector of covariates \( X \in \mathcal{X} \subset \mathbb{R}^{d_X} \). The observed data consists of a random sample \( (Y_i, T_i, Z_i, X_i), 1 \leq i \leq n \), from the population \( (Y, T, Z, X) \).

The following two assumptions define the role of the instrument Z.

**Assumption 1 (Conditional Independence).** \( \{(Y_i : t \in T), (T_z : z \in Z)\} \perp Z \mid X \).

**Assumption 2 (Common Support).** The support of \( X \) does not vary with \( Z \), that is, \( \text{supp}(X|Z = z) = \text{supp}(X) \) for every \( z \in Z \).

Assumption 1 states that the instrument \( Z \) is independent of the potential treatments and outcomes when taking into account the covariates \( X \). Assumption 2 assumes that for any value of the instrument, the same set of covariate values can be observed.

We use the Oregon Health Insurance Experiment (Finkelstein et al. 2012) as a running example to illustrate the concepts.\(^4\) The instrument \( Z = 0,1 \) is a random lottery that determines whether a participant is awarded the option to enroll in a Medicaid program. The treatment \( T \) is the actual insurance plan chosen by the individual, which falls into one of three categories: Medicaid \( (m) \), non-Medicaid insurance plans \( (nm) \), and no health insurance \( (no) \). Thus, \( T = \{m, nm, no\} \). The potential treatments \( T_0 \) and \( T_1 \) represent the insurance plan that the participant would have selected if the lottery status were set at \( Z = 0 \) and \( Z = 1 \), respectively. The outcome of interest, \( Y \), is a measure of health status. The potential outcomes \( Y_m, Y_{nm}, \) and \( Y_{no} \) represent the potential health status that would have been observed if the insurance plan were set to be \( m, nm, \) or \( no \), respectively.

As noted in the introduction, our model incorporates conditioning covariates, a feature not explicitly considered in the article by Heckman and Pinto (2018).\(^5\) In the empirical study, the covariates consist of (a) control variables that are designed to correlate with the lottery probability, such as household size and survey wave, and (b) variables that could potentially enhance power and efficiency, such as demographic factors and pre-randomization measurements of health outcomes.\(^6\)

Most empirical studies consider incorporating the above two types of conditioning covariates. In observational studies, researchers include additional covariates in the analysis to ensure the instrument’s validity. Even in experimental studies such as Finkelstein et al. (2012), where the instrument is guaranteed to be valid conditioning on a small set of covariates, there is still a motivation to include more covariates to enhance the efficiency of estimators and the statistical power of tests. Hence, it is crucial to study the implications of including a large number of covariates or even high-dimensional covariates in the econometric procedure.\(^7\)

The description above outlines a random sampling model in which the researcher only observes one potential health outcome associated with the actual insurance plan. This means that the sample of \( Y \), observed for an individual with insurance plan \( T = t \), comes from the conditional distribution of \( Y_t \) given \( T = t \) rather than from the marginal distribution of \( Y_t \). In general, this creates identification issues and poses challenges for causal inference. To address these problems, we introduce additional structures to the model.

**Assumption 3 (Unordered Monotonicity).** For any \( t \in T, z, z’ \in Z \), either

\[ P(1[T_z = t]) \geq P(1[T_{z’} = t] | X) = 1, \text{ or} \]

\[ P(1[T_z = t]) \leq P(1[T_{z’} = t] | X) = 1. \]

Assumption 3 represents the conditional version of the unordered monotonicity condition proposed by Heckman and Pinto (2018). This assumption implies that, when focusing

---

\(^4\)A more detailed background on the experiment is postponed to Section 6.

\(^5\)Heckman and Pinto (2018) focus on the identification analysis while keeping the conditioning covariates implicit. In other words, their analysis can be interpreted as being conditional on \( X = x \) for a specific value of \( x \). For the estimation issues addressed in our article, however, the role of covariates is more prominent. For example, the efficiency bound is computed directly through conditional moments of various variables given \( X \); and the smoothness of these conditional moments and the dimension of \( X \) all play a role in determining whether the causal parameter can be \( \sqrt{n} \)-estimated.

\(^6\)These two sets of conditioning covariates are considered in the original paper by Finkelstein et al. (2012).

\(^7\)In this article, we consider the set of covariates as given by the researcher’s field knowledge or data availability. The issue of model specification, including the selection of covariates, is beyond the scope of our study.
on a specific treatment level \( t \) and a pair \((z, z')\) of instrument values, the binary environment should adhere to the standard monotonicity constraint found in the LATE model. Specifically, the unordered monotonicity condition mandates that a change in the instrument uniformly moves all agents toward or away from each possible treatment value. The combination of Assumptions 1–3 yields a multivalued treatment equivalent from each possible treatment value. The combination of the unordered monotonicity condition found in the LATE model. Specifically, there are only five distinct types of insurance choice behaviors, defined by the pair of potential insurance plans \((T_0, T_1)\). These five types are listed in Table 1, where the third and fourth columns of the table are explained later.

The two groups of never-takers choose not to enroll in Medicaid regardless of the offer. Always-takers manage to enroll even without winning the lottery, presumably through mechanisms explained in Finkelstein et al. (2012) on p. 1080. The no- and mm-compliers transition to Medicaid from no insurance and non-Medicaid insurance, respectively, when such an opportunity appears beneficial because it can be directly used in the identification results.

As a general notation, for an integer \( k = 0, \ldots, N_Z \), we use \( \Sigma_{t,k} \) to represent the set of types in which the treatment level \( t \) appears exactly \( k \) times. Mathematically, this can be written as \( \Sigma_{t,k} = \{ s \in S : \sum_{i=1}^{N_Z} 1[s[i] = t] = k \} \), where \( s[i] \) denotes the \( i \)th element of the vector \( s \). The collection \( \Sigma_{t,k} \) contains a partition of \( S \). In the fourth column of Table 1, we list the type set notations for each insurance choice behavior type. Notice that the type set need not be a singleton though. For example, \( \Sigma_{m,1} = \{s_4, s_5\} \) contains no-compliers and mm-compliers.

### 2.2. Identification of Causal Parameters

For individuals with type \( S \) belonging to the same type set \( \Sigma_{t,k} \), their treatment response is in a way homogeneous. As a result, it becomes easier to identify the marginal distribution of the potential outcome \( Y_t \) within each \( \Sigma_{t,k} \). More specifically, we define the local average structural functions (LASF) and the local average structural functions for the treated (LASFT) as follows:

\[
\text{LASF}: \beta_{t,k} = \mathbb{E}[Y_t \mid S \in \Sigma_{t,k}], \\
\text{LASFT}: \gamma_{t,k} = \mathbb{E}[Y_t \mid S \in \Sigma_{t,k}, T = t].
\]

These parameters serve as natural extensions of the causal parameters examined in the binary LATE model.

To identify the LASF \( \beta_{t,k} \), we first need to determine \( p_{t,k} = \mathbb{P}(S \in \Sigma_{t,k}) \), the size of the subpopulation on which \( \beta_{t,k} \) is defined. Although \( S \) is unobserved, we can still deduce the probability \( p_{t,k} \) by linearly combining treatment probabilities given each instrument level. For instance, from the insurance choice behaviors listed in Table 1, we can infer that the size of no-never-takers is identified by the size of the subpopulation of individuals who do not have any insurance coverage even when given the opportunity to enroll in Medicaid. Subsequently, the size of no-compliers is identified by subtracting the size of no-never-takers from the subpopulation of individuals who do not have any insurance coverage without the opportunity to enroll in Medicaid. Following the same reasoning, we can identify the LASF by linearly combining conditional expectations of outcomes given each instrument level.

To formally generalize the above discussion and obtain identification results in the GLATE model, we introduce some matrix notations related to the type \( S \) that represent the linear operation mentioned in the previous paragraph. For each treatment level \( t \in \mathcal{T} \), let \( B_t \) be a binary matrix of the same dimension as the response matrix \( R \) with each element of \( B_t \) signifying whether the corresponding element in the response matrix is

### Table 1. Types of insurance choice behaviors.

| Type name         | Potential insurance plans \((T_0, T_1)\) | \(S\) | Type set \(\Sigma_{t,k}\) |
|-------------------|------------------------------------------|------|---------------------------|
| no-never-takers   | \((no, no)\)                              | \(s_1\) | \(\Sigma_{no,2}\)        |
| mm-never-takers   | \((nm, nm)\)                              | \(s_2\) | \(\Sigma_{nm,2}\)        |
| always-takers     | \((m, m)\)                                | \(s_3\) | \(\Sigma_{m,2}\)         |
| no-compliers      | \((no, m)\)                               | \(s_4\) | \(\Sigma_{no,1}\)        |
| mm-compliers      | \((nm, m)\)                               | \(s_5\) | \(\Sigma_{nm,1}\)        |

\(N_S = 5\), and the response matrix is

\[
R = (s_1, s_2, s_3, s_4, s_5) = \begin{pmatrix} no & nm & m & no & nm \\ no & nm & m & m & m \end{pmatrix}.
\]
t. In other words, $B_{t[j, f]}$, the $(j, f)$th element of $B_t$, is whether $T_{ji}$ equals $t$ for the subpopulation $S = s_j$. Define $b_{t,k} = \{1|s_i \in \Sigma_{t,k}\} \cdot B_{t,k}^+$, where $B_{t,k}^+$ is the Moore-Penrose inverse of $B_t$.  

For convenience, we also need some notations regarding conditional expectations. Let $\pi(X) = (\pi_{t}(X), \ldots, \pi_{T(2)}(X))$ be the vector of functions that describes the conditional distribution of the instrument $X$. For each treatment level $t \in T$, let $P_t(X) = (P_{t,1}(X), \ldots, P_{t,2}(X))$ with $P_{t,z}(X) = \mathbb{P}(T = t | Z = z, X)$ be the vector that describes the conditional treatment probabilities given each level of the instrument. Denote $Q_{t}(X) = (Q_{1,1}(X), Q_{1,2}(X))$ with $Q_{1,z}(X) = \mathbb{E}[Y|T = t | Z = z, X]$ as the vector containing the conditional outcomes for each treatment level $t$. Notice that the functions $\pi, P_t, Q_t$ are calculated from the observed distribution of $(Y, T, Z, X)$ and are thus identified.

**Theorem 2.1 (Identification of LASF).** Let Assumptions 1–3 hold. For $t \in T$ and $k \in 1, \ldots, N_Z$, the following identification results hold.

(i) The type set probability is identified by $p_{t,k} = \mathbb{P}(S = \Sigma_{t,k}) = b_{t,k}E[P_t(X)]$.

(ii) If $p_{t,k} > 0$, the LASF is identified by $\beta_{t,k} = b_{t,k}E[Q_t(X)] / p_{t,k}$.

**Theorem 2.1** identifies $p_{t,k}$, the size of the subpopulation $\Sigma_{t,k}$, and the mean potential outcome for that subpopulation. The identification fails only when the type set is $\Sigma_{t,0}$, in which case the individual never chooses the treatment $t$. Since the conditional distribution of $X$ given $S$ is not directly observable, identifying the "unconditional-on-covariates" LASF is not a straightforward extension from the conditional identification results in Heckman and Pinto (2018). Similar to the approaches employed in Abadie (2003) and Frölich (2007), we apply Bayes’ rule to achieve the identification of unconditional LASF under the "conditional-on-covariate" validity of instruments (Assumption 1).

The subsequent theorem presents the identification result for the LASF, which constitutes an additional contribution to the identification results, building upon Heckman and Pinto (2018). Let $Z_{t,k} \subset Z$ be the set of instrument values that induce the treatment level $t$ in the type set $\Sigma_{t,k}$. To clarify, $Z_{t,k} = \{z_i | Z: s[i] = t, \text{for all } s \in \Sigma_{t,k}\}$, where $s[i]$ denotes the $i$th element of the vector $s$. We then define $\pi_{t,k} = \sum_{s \in Z_{t,k}} \pi_{s}$ as the total probability of those instrument values.

**Theorem 2.2 (Identification of LASFT).** Let Assumptions 1–3 hold. For $t \in T$ and $k \in 1, \ldots, N_Z$, $Z_{t,k}$ is nonempty, and the following identification results hold.

(i) The treatment probability within a type set is identified by $q_{t,k} = \mathbb{P}(T = t, S = \Sigma_{t,k}) = b_{t,k}E[P_t(X)\pi_{t,k}(X)]$.

(ii) If $q_{t,k} > 0$, then the LASFT is identified by $\gamma_{t,k} = b_{t,k}E[Q_t(X)\pi_{t,k}(X)] / q_{t,k}$.

---

6The Moore-Penrose inverse $A^+$ of a matrix $A$ is the unique matrix that satisfies $AA^+A = A$, $A^+AA^+ = A^+$, and that both $AA^+$ and $A^+A$ are symmetric.

---

To clarify the notation used, the subscript 1 in $p_{n,0,1}$ and $\beta_{n,0,1}$ represents $k = 1$, whereas the subscript 1 in $p_{n,0,1}$ and $Q_{n,0,1}$ corresponds to $z = 1$.
The two denominators in the above expressions are both equal to the type probability of compliers. Then the usual identification of the LATE parameter (e.g., Theorem 1 of Frölich 2007) follows:

\[ \mathbb{E}[Y_1 - Y_0 \mid T_1 > T_0] = \frac{\int (\mathbb{E}[Y \mid Z = 1, X = x] - \mathbb{E}[Y \mid Z = 0, X = x])f_X(x)dx}{\int (\mathbb{E}[T \mid Z = 1, X = x] - \mathbb{E}[T \mid Z = 0, X = x])f_X(x)dx}, \]

where \( f_X \) represents the marginal density function of \( X \).

### 3. Semiparametric Efficiency

In this section, we compute the efficient influence function (EIF) and semiparametric efficiency bound (SPEB) for the LASF and LASFT parameters. In Appendix C of the Supplementary Material, we examine general parameters that are implicitly defined through moment restrictions.

#### 3.1. Efficiency Bound for LASF and LASFT

For the rest of the article, we assume that \( Y_t, t \in \mathcal{T} \) have finite second moments. This is necessary since we are studying efficiency. Let \( t \) denote the column vector of ones and \( \zeta(Z, X, \pi) \) the diagonal matrix with the diagonal elements being \( 1[Z = z]/\pi(Z, x), z \in \mathcal{Z} \). The following theorem gives the efficient influence function (EIF) and the SPEB for the parameters identified in the preceding section.

**Theorem 3.1 (SPEB for LASF and LASFT).** Let Assumptions 1–3 hold. Let \( t \in \mathcal{T} \) and \( k \in \{1, \ldots, N_Z\} \). Assume that \( p_{t,k}, q_{t,k} > 0 \).

(i) The semiparametric efficiency bound for \( \beta_{t,k} \) is given by the variance of the efficient influence function

\[
\psi^{\beta_{t,k}}(Y, T, Z, X, \beta_{t,k}, p_{t,k}, Q_t, P_t, \pi) = \frac{1}{p_{t,k}} b_{t,k} \left( \zeta(Z, X, \pi) \left( (Y 1(T = t) - Q_t(X)) + Q_t(X) \right) \right) - \frac{\beta_{t,k}}{p_{t,k}} b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 - T) - P_t(X) \right) + P_t(X) \right). \tag{1}
\]

(ii) The semiparametric efficiency bound for \( \gamma_{t,k} \) is given by the variance of the efficient influence function

\[
\psi^{\gamma_{t,k}}(Y, T, Z, X, \gamma_{t,k}, q_{t,k}, Q_t, P_t, \pi) = \frac{1}{q_{t,k}} b_{t,k} \left( \zeta(Z, X, \pi) \left( (Y 1(T = t) - Q_t(X)) \pi_{t,k}(X) \right) + Q_t(X) 1(Z \in \mathcal{Z}_{t,k}) \right) - \frac{\gamma_{t,k}}{q_{t,k}} b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 - T) - P_t(X) \right) \pi_{t,k}(X) \right) + P_t(X) 1(Z \in \mathcal{Z}_{t,k}) \right). \tag{2}
\]

(iii) The semiparametric efficiency bound for \( p_{t,k} \) is given by the variance of the efficient influence function

\[
\psi^{p_{t,k}}(T, Z, X, p_{t,k}, P_t, \pi) = b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 \{T = t\} - P_t(X)) \right) - P_t(X) + P_t(X) \right) - p_{t,k}. \tag{3}
\]

(iv) The semiparametric efficiency bound for \( q_{t,k} \) is given by the variance of the efficient influence function

\[
\psi^{q_{t,k}}(T, Z, X, q_{t,k}, P_t, \pi) = b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 \{T = t\} - P_t(X)) \right) + P_t(X) 1(Z \in \mathcal{Z}_{t,k}) \right) - q_{t,k}. \tag{4}
\]

In the binary LATE model, the first two parts of Theorem 3.1 reduce to Theorem 2 of Frölich (2007) and Theorem 2 of Hong and Nekipelov (2010). If we maintain the multiplicity of treatment and instrument but assume unconfoundedness by having \( T = Z \), then the result reduces to Theorem 1 of Cattaneo (2010).

The EIF in Theorem 3.1 can be interpreted as the moment condition from the identification results modified by an adjustment term due to the presence of unknown infinite-dimensional parameters. Take \( \psi^{\beta_{t,k}} \) as an example, the terms

\[
b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 \{T = t\} - Q_t(X)) \right) \right) / p_{t,k}, \quad \text{and} \quad \beta_{t,k} b_{t,k} \left( \zeta(Z, X, \pi) \left( (1 \{T = t\} - P_t(X)) \right) \right) / p_{t,k}
\]

are respectively the adjustment terms due to the presence of \( Q_t \) and \( P_t \).

From the expression of \( \psi^{\beta_{t,k}} \), we can see that the SPEB would be large when \( p_{t,k} \) is small. This is because \( p_{t,k} \) measures the size of the subpopulation \( S \in \mathcal{S}_{t,k} \) on which the LASF is estimated. When \( p_{t,k} \) is small, we could run into the weak identification issue. In Section 5, we study inference procedures that are robust against weak identification issues.

One benefit of the EIFs is that we can easily calculate the covariance matrix of different estimators. Consider an example where we are interested in two LASFs \( \beta_1 \) and \( \beta_2 \), whose EIF is given by \( \psi_1 \) and \( \psi_2 \), respectively. If the two estimators \( \beta_1 \) and \( \beta_2 \) are both semiparametric efficient, then their covariance matrix equals \( \mathbb{E}[\psi_1 \psi_2^T] \).

#### 3.2. Efficient Estimation

The derived SPEB helps determine whether an estimation procedure is efficient. In this section, we focus on the conditional expectation projection (CEP) estimator. Define \( h_{Y,T,Z}(X) = \mathbb{E} \left[ (Z \mid Z = z) Y 1\{T = t \mid X \} \right] \) and \( h_{T,Z}(X) = \mathbb{E} \left[ (Z \mid Z = z) 1\{T = t \} \right] \). The CEP procedure first estimates \( \pi_{t,k}, h_{Y,T,Z} \), and \( h_{T,Z} \) by using nonparametric estimators \( \hat{\pi}_z, \hat{h}_{Y,T,z} \), and \( \hat{h}_{T,z} \), respectively. These estimators can be constructed based on series or local polynomial estimation. The nuisance functions \( Q_{t,z} \) and \( P_{t,z} \) are estimated using \( \hat{Q}_{t,z} = \hat{h}_{Y,T,z}/\hat{\pi}_z \).

---

10 As with the proofs provided in Frölich (2007) and Hong and Nekipelov (2010), our demonstration of the efficiency bound does not involve the explicit construction of parametric submodels that satisfy the model's identification assumptions. This method, as emphasized in the recent study by Navjeevan, Pinto, and Santos (2023), could result in a loose efficiency bound when the model is locally overidentified, as defined by Chen and Santos (2018). However, similar to the binary case, the GLATE model that includes multivalued treatment is locally just-identified, as defined by Chen and Santos (2018). Consequently, our efficiency bound is accurate. We appreciate the editor's suggestion on this issue.

11 The terminology “conditional expectation projection” is adopted from Chen, Hong, and Tiarozzi (2008) and Hong and Nekipelov (2010), whereas Hahn (1998) refers to these estimators as “nonparametric imputation based estimators.”
and \( \hat{P}_{t,z} = \hat{h}_{t,z} / \hat{\pi}_z \). The vectors of estimators \( \hat{Q}_t \), \( \hat{P}_t \), and \( \hat{\pi} \) are stacked in an obvious way. Let \( \hat{\pi}_{t,k} = \sum_{z \in Z_t} \hat{\pi}_z \). The CEP estimators for the structural parameters are defined by

\[
\hat{P}_{t,k} = \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{P}_i(X_i), \quad \hat{Q}_{t,k} = \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{Q}_i(X_i), \quad \hat{\pi}_{t,k} = \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{\pi}_i(X_i).
\]

The next proposition shows that the CEP estimators are semiparametrically efficient. The result is similar in style to Hahn's (1998) Proposition 4 that the low-level regularity conditions are omitted. Instead, the proposition assumes the high-level condition that the CEP estimators are asymptotically linear, which means they are asymptotically equivalent to sample averages. More formally, an estimator \( \hat{\beta} \) of \( \beta \) is asymptotically linear if it admits an influence function. That is, there exists a function \( \Psi \) with zero mean and finite variance such that \( \sqrt{n}(\hat{\beta} - \beta) = \sum_{i=1}^{n} \Psi(Y_i, T_i, Z_i, X_i) / \sqrt{n} + o_p(1) \).

**Proposition 3.2.** Suppose the CEP estimators are asymptotically linear, then they achieve the semiparametric efficiency bound.

The reason that this type of estimator is efficient is well explained in Ackerberg et al. (2014). The estimation problem here falls into their general semiparametric model, where the finite-dimensional parameter of interest is defined by unconditional moment restrictions. They show that the semiparametric two-step optimally weighted GMM estimators, the CEP estimators in this case, achieve the efficiency bound since the parameters of interest are exactly identified. Discussions related to this phenomenon can also be found in Chen and Santos (2018).

To be more concrete, we construct CEP estimators that use local polynomial regression in the first step. Let \( K \) be a kernel function on \( \mathbb{R}^{d_k} \) and \( a = a_n \) a bandwidth sequence. For \( x \in X \), define \( r_j(X_i - x) \) as the vector containing polynomials of \( X_i - x \) with degree from 0 to \( \lambda: r_j(X_i - x) = (1, X_i - x_1, \ldots, X_i - x_{d_k}, \ldots, (X_i - x_1)^j, \ldots, (X_i - x_{d_k})^j)' \). The estimator \( \hat{h}_{Y,t,z}(x) \) is the intercept coefficient obtained from the weighted least squares regression of \( I(Z_i = z)\frac{1}{1(1)} \) on \( r_j(X_i - x) \) with weights \( K((x - X_i)/a) \). The estimators \( \hat{h}_{t,z}(x) \) and \( \hat{\pi}_z(x) \) are defined analogously by replacing the regressand with \( I(Z_i = z)\frac{1}{1(1)} \) and \( I(Z_i = z) \), respectively.

In the following theorem, we adopt the assumptions on local polynomial regression from Theorem 4 in Frölich (2007) and Theorem 3 in Heckman, Ichimura, and Todd (1998) and demonstrate that the corresponding CEP estimator is efficient. Specifically, it defines the range of the tuning parameter \( a \) that works; \( a \) needs to converge faster than \( n^{-\lambda / 2d_k} \) but slower than \( (n/\log n)^{-1/d_k} \).

**Theorem 3.3.** Assume that the functions \( h_{Y,t,z}, h_{t,z}, \) and \( \pi_z \) are \( \lambda \)-times continuously differentiable with Hölder continuous \( \lambda \)th derivatives, where \( \lambda > d_k \); the propensity score \( \pi_z \) is bounded away from zero; the bandwidth \( a \) satisfies \( na^{2d_k} / \log n \to \infty \) and \( na^{2k} \to 0 \); and the kernel function \( K \) is symmetric, compactly supported, Lipschitz continuous, and has moments of order 1 through \( \lambda \) equal to zero. Under these assumptions, the CEP estimators, which use local polynomial regression of order \( \lambda \) in the first step, are asymptotically linear and thus achieve the semiparametric efficiency bound.

### 3.3. Other Policy-Relevant Parameters

We next examine the efficient estimation of other policy-relevant parameters that can be derived from the parameters \( (\beta_{t,k}, \gamma_{t,k}, \pi_{t,k}, \eta_{t,k}) \). As an example, consider the type set \( \Sigma_t = U_{k=1}^{N_{Zt}} \Sigma_{t,k} \) which is referred to as \( t \)-switchers. This subpopulation contains individuals who switch between \( t \) and other treatments when given different levels of instruments. It is a generalization of the concept of compliers in the binary LATE framework. The LASF for the subpopulation \( \Sigma_t \) is defined by \( \beta_t = \mathbb{E} [Y_i | \pi_t, t, S \in \Sigma_t] = \sum_{k=1}^{N_{Zt}} \beta_{t,k} \pi_{t,k} / \sum_{k=1}^{N_{Zt}} \pi_{t,k} \). Similarly, one can also define \( \gamma_t = \mathbb{E} [Y_i | \pi_t, T = t, S \in \Sigma_t] = \sum_{k=1}^{N_{Zt}} \gamma_{t,k} \pi_{t,k} / \sum_{k=1}^{N_{Zt}} \pi_{t,k} \), which represents the LASFT for the subpopulation of \( t \)-treated \( t \)-switchers. In the context of our empirical study, the group designated as \( m \)-switchers encompasses both \( no \) - and \( nm \)-compliers. This group aligns with the compliers as defined in the binary LATE model.

More generally, let \( \phi = \phi(p, q, \beta, \gamma) \) be a finite-dimensional parameter, where \( \phi(\cdot) \) is a known continuously differentiable function, and \( p \) is the vector containing all identifiable \( p_{t,k} \)'s, that is, \( p = \{p_{t,k} : t \in T, 1 \leq k \leq N_{Zt}\} \). Let \( q, \beta, \) and \( \gamma \) be defined analogously. A natural estimator can be defined through the CEP estimates, \( \hat{\phi}(\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}) \). The delta-method can help calculate the efficiency bound of \( \phi \) and show the efficiency of \( \hat{\phi}(\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}) \). In fact, by Theorem 25.47 of van der Vaart (1998), we immediately have the following corollary, which shows that plug-in estimators are efficient.

**Corollary 3.4.** The semiparametric efficiency bound of \( \phi \) is given by the variance of efficient influence function \( \psi' = \sum_{p \in \mathcal{P}} \frac{\partial \phi}{\partial p} \psi + \sum_{q \in \mathcal{Q}} \frac{\partial \phi}{\partial q} \psi' + \sum_{\beta \in \mathcal{B}} \frac{\partial \phi}{\partial \beta} \psi \beta + \sum_{\gamma \in \mathcal{G}} \frac{\partial \phi}{\partial \gamma} \psi' \gamma \), where the partial derivatives are evaluated at the true parameter value. Moreover, the plug-in estimator \( \hat{\phi}(\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}) \), based on the CEP estimators \( \hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma} \), achieves the efficiency bound.

### 4. Robustness

In the preceding section, the EIF was employed as a device for calculating the SPEB. In the current section, we directly use the EIF to formulate moment conditions. These moment conditions are appealing as they ensure both double robustness and Neyman orthogonality, the two primary focuses of this section.

For clarity in notation, throughout the remainder of the article, we use a superscript \( o \) to denote the true value when necessary. For example, in contexts where both \( \pi^o \) and \( \pi \) are present, the former represents the true instrument probability while the latter signifies a generic function.

### 4.1. Double Robustness

We focus on the LASF \( \beta_{t,k} \). The same analysis can be applied to the other parameters. To avoid notational burden in the main
text, we drop the subscript \((t, k)\) in \(\beta_{t,k}, P_{t,k},\) and \(b_{t,k}\), and the subscript \(t\) in \(P_t,\) and \(Q_t.\) It is straightforward to verify that the EIF \(\psi^\beta\) has zero mean. However, we do not want to use \(\psi^\beta\) itself as the estimating equation since it contains 1-dimensional nuisance parameters. To deal with this problem, we simply multiply \(\psi^\beta\) by \(p\) and define
\[
\psi(Y, T, Z, X, \beta, Q, P, \pi) = p\psi^\beta(Y, T, Z, X, \beta, Q, P, \pi)
\]
\[= b(\zeta(Z, X, \pi) (1(Y1(T = t)) - Q(X)) + Q(X)) - \beta b(\zeta(Z, X, \pi) (1(T = t) - P(X)) + P(X)).
\]
The corresponding moment condition is
\[
E[\psi(Y, T, Z, X, \beta^o, Q^o, P^o, \pi^o)] = 0. \tag{2}
\]
This moment condition is doubly robust, as demonstrated in the following proposition.

**Proposition 4.1 (Double Robustness).** For an arbitrary vector of functions \((Q, P, \pi),\) we have \(E[\psi(Y, T, Z, X, \beta^o, Q^o, P^o, \pi)] = 0,\) and \(E[\psi(Y, T, Z, X, \beta^o, Q, P, \pi^o)] = 0.\)

The above proposition divides the nonparametric nuisance parameters into two groups, \(\pi\) and \((Q, P).\) The doubly robust moment condition is valid if either of these two groups of nuisance parameters is true. If the researcher uses parametric models for these nuisance parameters, then the structural parameter \(\beta\) can be recovered provided that at least one of the working nuisance models is correctly specified. Therefore, the doubly robust moment condition is “less demanding” on the researcher’s ability to devise a correctly specified model for the nuisance parameters.

### 4.2. Neyman Orthogonality

The second robustness property is Neyman orthogonality. Moment conditions with this property have reduced sensitivity with respect to the nuisance parameters. Formally, Neyman orthogonality means that the moment condition has zero Gateaux derivative with respect to the nuisance parameters. The result is presented in the following proposition.

**Proposition 4.2 (Neyman Orthogonality).** Let \((Q, P, \pi)\) be an arbitrary set of functions. For \(r \in [0, 1),\) define \(Q^r = Q^o + r(Q - Q^o),\) \(P^r = P^o + r(P - P^o),\) and \(\pi^r = \pi^o + r(\pi - \pi^o).\) Suppose that \(sup_{r \in [0, 1]} \left| \frac{\partial}{\partial r} E[\psi(Y, T, Z, X, \beta^o, Q^r, P^r, \pi^r)] \right| < \infty,\) and \(\beta\) does not need to be the true parameter value.

In many econometric models, double robustness and Neyman orthogonality come in pairs. Discussions about their general relationships can be found in Chernozhukov et al. (2022d). In practice, double robustness is often used for parametric estimation, as previously explained, whereas Neyman orthogonality is used in estimation with the presence of possibly high-dimensional nuisance parameters.

Next, we apply the double/debiased machine learning (DML) method developed by Chernozhukov et al. (2018) to the moment condition (2). This estimation method works even when the nuisance parameter space is complex enough that the traditional assumptions, for example, Donsker properties, are no longer valid. The implementation details are explained below.

The nuisance parameters \(Q, P,\) and \(\pi\) are estimated using a cross-fitting method: Take an \(L\)-fold random partition of the data such that the size of each fold is \(n/L.\) For \(\ell = 1, \ldots, L,\) let \(I^\ell\) denote the set of observation indices in the \(\ell\)th fold and \(I^\ell = \bigcup_{\ell \neq i} I^i\) the set of observation indices not in the \(i\)th fold. Let \(\hat{Q}^\ell, \hat{P}^\ell,\) and \(\hat{\pi}^\ell\) be the nuisance parameter estimates constructed by using data from \(I^\ell,\) whose construction are explained later.

The DML estimator of \(\beta\) is constructed following the moment condition (2):\(^{14}\)

\[
\hat{\beta} = \frac{\left( \sum_{\ell=1}^{L} \sum_{i \in I^\ell} b(\zeta(Z_i, X_i, \hat{\pi}^\ell) (1(Y1(T_i = t)) - \hat{Q}^\ell(X_i)) + \hat{Q}^\ell(X_i)) \right)}{\left( \sum_{\ell=1}^{L} \sum_{i \in I^\ell} b(\zeta(Z_i, X_i, \hat{\pi}^\ell) (1(T_i = t)) - \hat{P}^\ell(X_i)) + \hat{P}^\ell(X_i)) \right)}. \tag{3}
\]

For the purposes of inference, it is necessary to estimate the asymptotic variance of \(\hat{\beta},\) denoted here as \(\sigma^2.\) The construction of such an estimator is made possible due to the fact that the moment condition for the DML estimator is formulated based on the EIF. As a result, the asymptotic variance \(\sigma^2\) equals the expectation of the squared EIF, expressed as \(\sigma^2 = E[\psi^\beta]^2 = E[\psi^2]^2.\) Importantly, all the components needed to assemble the estimator \(\hat{\sigma}^2\) have already been calculated during the estimation of \(\beta.\) The estimator of \(\sigma^2\), denoted \(\hat{\sigma},\) is simply the denominator of \(\hat{\beta}\) in (3):

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I^\ell} \left[ \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I^\ell} b(\zeta(Z_i, X_i, \hat{\pi}^\ell)) (1(Y1(T_i = t)) - \hat{P}^\ell(X_i)) + \hat{P}^\ell(X_i)) \right]^2 / \hat{\beta}^2. \tag{4}
\]

Then we can estimate the asymptotic variance by

\[
\hat{\beta}^2 = \frac{\left( \sum_{\ell=1}^{L} \sum_{i \in I^\ell} b(\zeta(Z_i, X_i, \hat{\pi}^\ell) (1(Y1(T_i = t)) - \hat{Q}^\ell(X_i)) + \hat{Q}^\ell(X_i)) \right)}{\left( \sum_{\ell=1}^{L} \sum_{i \in I^\ell} b(\zeta(Z_i, X_i, \hat{\pi}^\ell) (1(T_i = t)) - \hat{P}^\ell(X_i)) + \hat{P}^\ell(X_i)) \right)}. \tag{3}
\]

To better illustrate the interconnected calculation of the three estimators \(\hat{\beta}, \hat{\sigma},\) and \(\hat{\pi}\) and to facilitate their coding implementation, we define \(\hat{\pi}^{n, L} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I^\ell} (1(T_i = t))\) and the following terms:

\[
\omega_{k, t}^1 = b(\zeta(Z_t, X_t, \hat{\pi}^\ell)) (1(Y1(T_t = t)) - \hat{Q}^\ell(X_t)) + \hat{Q}^\ell(X_t),
\]

\[^{13}\text{The full subscripts are kept in the Appendices that contain proofs.}\]

\[^{14}\text{In two-step semiparametric estimations, Donsker properties are usually required so that a suitable stochastic equicontinuity condition is satisfied. See, for example, Assumption 2.5 in Chen, Linton, and van Keilegom (2003).}\]

\[^{15}\text{We study the DML2 estimator defined in Chernozhukov et al. (2018). Another estimator, the DML1 estimator, is also proposed in that same paper. We do not study the DML1 estimator since it is asymptotically equivalent to DML2, and the authors generally recommend DML2.}\]
\(\omega_{t,\ell}^2 = b(\xi(Z_i, X_i, \pi^\ell)) (1(T_i = t) - \hat{\beta}^\ell(X_i)) + \tilde{\beta}^\ell(X_i).\)

Consequently, we can express the three estimator as 
\(\hat{\beta} = \frac{\hat{E}_{n,l}[\omega_{t,\ell}]}{\hat{E}_{n,l}[\omega_{t,\ell}^2]}, \hat{\beta} = \hat{E}_{n,l}[\omega_{t,\ell}],\) and \(\hat{\sigma} = \sqrt{\hat{E}_{n,l}[\omega_{t,\ell}^2]} / \hat{E}_{n,l}[\omega_{t,\ell}^2].\)

We now turn to the task of establishing convergence results for the DML estimator, which are uniform across a class of data generating processes (DGP), defined as follows. For any two constants \(c_1 > c_0 > 0,\) let \(P(c_1, c_0)\) be the set of joint distributions of \((Y, T, Z, X)\) such that (i) \(p \in [c_0, 1],\) and (ii) \(E[\psi(Y)] \geq c_0,\) and \(|Y| |T = t|, |Y| |T = t| - Q_0^\ell|X| < c_1.\) The first condition of \(P(c_1, c_0)\) excludes the case where \(\beta\) is weakly identified (specifically when \(p\) can become arbitrarily close to 0). The following theorem states the asymptotic properties of the DML estimator. Specifically, confidence regions can be constructed by inverting standard \(t\)-tests.

**Theorem 4.3.** Let Assumptions 1–3 hold. Assume the following conditions on the nuisance parameter estimators \((\hat{Q}^\ell, \hat{P}^\ell, \hat{\pi}^\ell):\)

(i) For \(z \in Z, |\hat{Q}^\ell|\) is bounded, \(\hat{P}^\ell\) and \(\hat{\pi}^\ell\) are bounded away from zero.

(ii) \(\max_{z \in Z} \left(\|\hat{Q} - Q\|_2 \vee \|\hat{P} - P\|_2 \vee \|\hat{\pi} - \pi\|_2\right) = o_p(n^{-1/4}).\)

Then the estimator \(\hat{\beta}\) obeys that \(\sigma^{-1}\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0,1)\) uniformly over the DGP in \(P(c_0, c_1).\) Moreover, the above convergence result continues to hold when \(\sigma\) is replaced by the estimator \(\hat{\sigma}\).

In low-dimensional settings, the uniform convergence rate for the first-step estimators of the nuisance parameters can be achieved using standard methods. For example, minimax optimal uniform convergence rates are given in Stone (1982) subject to different smoothness conditions. For specific nonparametric procedures, refer to Hansen (2008) and Masry (1996) for local polynomial estimators and Chen and Christensen (2015) for series estimators. In high-dimensional settings, several machine learning methods are explored in Chernozhukov et al. (2018), including lasso, tree-based methods, neural networks, and their ensembles. Their findings indicate that the choice of machine learning method for estimating nuisance parameters does not lead to qualitative differences, and various methods yield broadly consistent results across multiple empirical studies.

We present an implementation procedure tailored for the multiple discrete treatment and instrument choices in the GLATE model, which can be readily implemented using the R package glmnet (v4.1-3; Friedman, Tibshirani, and Hastie 2010; Simon et al. 2011; Tay, Narasimhan, and Hastie 2023). To estimate the conditional probabilities \(\pi_Z\) and \(P_{t,x},\) we employ multinomial regressions with an elastic net penalty. Define \(\hat{\pi}_Z(X_i) = \exp(\hat{\mu}_{\hat{P}}(\hat{\mu}_{\hat{Q}} + (\hat{\mu}_{\hat{Q}}^\top X_i)) / \sum_{z \in Z} \exp(\hat{\mu}_{\hat{Q}} + (\hat{\mu}_{\hat{Q}}^\top X_i)),\) where the coefficients \((\hat{\mu}_{\hat{Q}}, \hat{\mu}_{\hat{Q}}^\top): z \in Z\) are obtained by minimizing the following penalized (negative) log-likelihood function 
\([-\sum_{t \in T} \sum_{z \in Z} 1(Z_i = z)(\mu_{0z} + \mu_{z}^\top X_i) - \log \left(\sum_{z' \in Z} \exp(\mu_{0z}^\ell + (\mu_{z}^\ell)^\top X_i)\right) + \alpha(1 - \alpha)\|\mu_{0z}^\ell\|_2^2 + \alpha\|\mu_{z}^\ell\|_2^2].\)

In this expression, \(\mu_{0z}^\ell\) is the \(d_X \times N_Z\) matrix that combines the column vectors \(\mu_{0z}^\ell, z \in Z.\) The norms \(\|\cdot\|_2^2\) and \(\|\cdot\|_2^2\) represent the Frobenius norm and the matrix norm that sums the absolute value of each entry, respectively. The elastic net penalty is controlled by \(\alpha \in [0,1],\) which bridges the gap between lasso regression (\(\alpha = 1\)) and ridge regression (\(\alpha = 0\)). The penalization parameter \(\lambda\) can be determined through cross-validation. Similarly, the coefficients in the conditional probability \(P_{t,x}\) can be estimated by minimizing 
\[-\sum_{t \in T} \sum_{z \in Z} 1(T_i = t)(\mu_{0z} + \mu_{z}^\top X_i) - \log \left(\sum_{z' \in Z} \exp(\mu_{0z} + (\mu_{z}^\ell)^\top X_i)\right) + \lambda \left[1 - \alpha\right]\|\mu_{0z}\|_2^2 + \alpha\|\mu_{z}^\top\|_2^2,\]

where \(\mu_{0z}^\ell\) is the \(d_X \times N_T\) matrix that combines the column vectors \(\mu_{0z}^\ell, t \in T.\) For estimating the conditional expectations \(Q_{t,x},\) we use standard linear regression with elastic net penalty. In Section 6, we showcase the application of this specific implementation procedure in the context of our empirical study.

### 5. Weak Identification

The convergence result established in Theorem 4.3 is uniform over the set of DGP with type probability \(p\) bounded away from zero. However, when \(p\) can be arbitrarily close to zero, the identification of \(\beta\) becomes weak, leading to distortion of the uniform size of the test and poor asymptotic approximation in finite-sample settings. This section addresses this weak identification issue and proposes an inference procedure that is robust against such problems.

We start with a heuristic illustration of the weak identification problem. For simplicity in notation, we define \(\nu = \beta p\) and

\[\hat{\nu} = \frac{1}{n} \sum_{i=1}^n \sum_{t \in T_i} b\left(\xi(Z_i, X_i, \pi^\ell) (1(Y_i|T_i = t)) - \hat{Q}^\ell(X_i) + \hat{\pi}^\ell(X_i)\right).\]

After a simple calculation, we can write \(\hat{\beta} - \beta = (\sqrt{n}(\hat{\nu} - \nu) - \beta(\sqrt{n}(\hat{\nu} - \beta)))/(\sqrt{n}(\hat{\nu} - \nu) + \sqrt{n}p)\).

In this expression, we can interpret the estimation errors \(\sqrt{n}(\hat{\nu} - \nu)\) and \(\sqrt{n}(\hat{\nu} - \beta)\) as noise, while the term \(\sqrt{n}p\) represents the signal. Under standard asymptotics where \(p > 0\) is fixed, the noise terms are bounded in probability, while the signal term \(\sqrt{n}p\) → ∞. Consequently, the signal dominates the noise, and the estimator \(\hat{\beta}\) is consistent.

However, under asymptotics with a drifting sequence \(p = p_n \rightarrow 0\) and \(\sqrt{n}p\) converging to a finite constant, the signal and noise have the same magnitude, leading to the inconsistency of \(\hat{\beta}\). This issue represents the weak identification problem. In the weak IV literature, the so-called concentration parameter is a common measure of identification strength. In our case, the concentration parameter is given by \(\sqrt{n}p\), where \(\sqrt{n}p\) → ∞ corresponds to strong identification, and identification is considered weak when the limit of \(\sqrt{n}p\) is finite.

Weak identification is a finite-sample concern, yet it is formulated within the asymptotic framework. The utilization of asymptotics under drifting sequences, as illustrated above, is not aimed at modeling DGP that change with the sample size.\(^{15}\)
Rather, it serves as a tool for detecting the absence of uniform convergence. In fact, achieving control over the uniform size of the test is essential for addressing weak identification problems (Imbens and Manski 2004; Mikusheva 2007; Andrews, Cheng, and Guggenberger 2020). Formally, the uniform size of a test is defined as the large sample limit of the supremum of the rejection probability under the null hypothesis, with the supremum taken over the nuisance parameter space. In the context of testing a null hypothesis on $\beta$ within the GLATE model, the supremum covers all values of $p > 0$. Consequently, a well-designed test should have a rejection probability under the null that converges uniformly to the nominal size over $p \in (0, 1]$. As demonstrated in the previous discussion, the uniform size cannot be achieved using the standard $t$-statistic $\sqrt{n} (\hat{\beta} - \beta) / \hat{\sigma}$. Nonetheless, this lack of uniform convergence does not contradict Theorem 4.3, in which the uniform convergence of $\hat{\beta}$ is established only after constraining $p$ to be bounded away from zero.

Inference procedures that are robust against weak identification can be constructed by directly incorporating the null hypothesis into the test statistic. One example is the well-known Anderson-Rubin (AR) statistic in the weak IV literature, which can be generalized to the GLATE model. We first consider testing the two-sided hypothesis $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$. To ensure the uniform size control of the test, it is necessary for the test statistic to exhibit uniform convergence over the parameter space where (1) $\beta = \beta_0$, and (2) $p$ can approach zero arbitrarily closely. A null-restricted $t$-statistic can be obtained as follows. Notice that when $p > 0$, $\beta = \beta_0$ is equivalent to

$$0 = \nu - \beta_0 p = E [\psi (Y, T, Z, X, \beta_0, Q, P, \pi)]. \tag{5}$$

Its estimate can be written as

$$\hat{\nu} - \beta_0 \hat{p} = (\hat{\nu} - \nu) - (\hat{\beta} - \beta_0) p.$$

Under the null hypothesis $\beta = \beta_0$, the above estimate does not depend on the concentration parameter $\sqrt{n} \hat{p}$ and consists only of the noise terms $\hat{\nu} - \nu$ and $\hat{\beta} - \beta$, whose uniform convergence can be established directly.

For implementation, this test statistic can be obtained as a straightforward application of the DML procedure to the moment condition (5). As a consequence of Proposition 4.2, the above moment condition satisfies the Neyman orthogonality condition regardless of the true value of $\beta$. More specifically, the null-restricted $t$-statistic is defined to be

$$\hat{\rho} = \sqrt{n} (\hat{\nu} - \beta_0 \hat{p}) / \hat{\sigma}_\psi,$$ \tag{6}

where $\hat{\sigma}_\psi^2 = \frac{1}{n} \sum_{t=1}^{T} \sum_{i \in \pi_t} \psi (Y_t, T_t, Z_t, X_t, \beta_0, \hat{Q}_t, \hat{P}_t, \hat{\pi}_t)^2$. The corresponding test of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ rejects for large values of $|\hat{\rho}|$.

The same methodology can be applied to testing one-sided hypothesis $H_0 : \beta \leq \beta_0$ versus $H_1 : \beta > \beta_0$. Under the null hypothesis, $(\beta - \beta_0) p$ is nonpositive, suggesting that the test should reject for large values of $\hat{\rho}$. Notice that this relies on knowing the sign of $p$ due to the GLATE model structure. This restriction on the sign of $p$ is similar to knowing the first-stage sign in the linear IV model, which is studied by Andrews and Armstrong (2017) in the context of unbiased estimation.

We now define the set of DGP’s that allows $p$ to be arbitrarily close to zero. For any two constants $c_1 > c_0 > 0$, let $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ be the set of joint distributions of $(Y, T, Z, X)$ such that (i) $p \in (0, 1]$, and (ii) $E[\psi^2, \pi(X) \geq c_0, z \in \mathcal{Z}, |Y_1(T = t)|, |Y_1(T = t) - Q_t(X)] \leq c_1$. The superscript “WI” denotes weak identification. For any $\beta' \in \mathbb{R}$, let $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ be the subset of $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ in which the true value of the parameter $\beta$ is $\beta'$. In particular, $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ denotes the subset where the null hypothesis is true. The difference between $\mathcal{P}(c_0, c_1)$ and $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ is that $\mathcal{P}_{\beta}^{\text{W}}(c_0, c_1)$ allows the type probability $\rho$ to be arbitrarily small, whereas the type probabilities in $\mathcal{P}(c_0, c_1)$ are uniformly bounded away from zero. Denote $\mathcal{N}_c$ as the $\nu$th quantile of the standard normal distribution. The following theorem establishes that the above testing procedures have uniformly correct size and are consistent.

**Theorem 5.1.** Suppose the conditions on the nuisance parameter estimates in Theorem 4.3 hold. Let $\alpha \in (0, 1)$ be the nominal size of the tests.

(i) The null-restricted test that rejects $H_0 : \beta = \beta_0$ in favor of $H_0 : \beta = \beta_0$ when $|\hat{\rho}| > \mathcal{N}_{1-\alpha}$ has (asymptotically) uniformly correct size and is consistent. That is,

$$\sup \{ \mathbb{P}_P (|\hat{\rho}| > \mathcal{N}_{1-\alpha}) : P \in \mathcal{P}_{\beta}^{\text{W}}(c_0, c_1) \} \to \alpha,$$

and

$$\mathbb{P}_P (|\hat{\rho}| > \mathcal{N}_{1-\alpha}) \to 1, P \in \mathcal{P}_{\beta}^{\text{W}}(c_0, c_1), \beta \neq \beta_0.$$ (ii) The null-restricted test that rejects $H_0 : \beta \leq \beta_0$ in favor of $H_0 : \beta > \beta_0$ when $\hat{\rho} > \mathcal{N}_{1-\alpha}$ has (asymptotically) uniformly correct size and is consistent. That is,

$$\sup \{ \mathbb{P}_P (\hat{\rho} > \mathcal{N}_{1-\alpha}) : P \in \mathcal{P}_{\beta}^{\text{W}}(c_0, c_1), \beta \leq \beta_0 \} \to \alpha,$$

and

$$\mathbb{P}_P (\hat{\rho} > \mathcal{N}_{1-\alpha}) \to 1, P \in \mathcal{P}_{\beta}^{\text{W}}(c_0, c_1), \beta > \beta_0.$$ The confidence region for $\beta$ can be constructed by inverting the corresponding two-sided test. The $1 - \alpha$ confidence region, $\mathcal{C}_{1-\alpha}$, contains all values of $\beta_0$ that cannot be rejected by the test: $\mathcal{C}_{1-\alpha} = \{ \beta : |\hat{\rho}(\beta_0)| \leq \mathcal{N}_{1-\alpha} \}$, where $|\hat{\rho}(\beta_0)|$ is defined by (7), and its dependence on $\rho$ is made explicit.

In Appendix B, we report the coverage results obtained from a simulation study that employs the data generating process studied by Hong and Nekipelov (2010), with modifications to address multiple treatments and weak identification issues. Specifically, we examine a data generating process where a type probability $p_n$ drifts to zero as $n$ increases. Our findings demonstrate that the weak-identification-robust confidence region, constructed for the weakly identified LASF, produces satisfactory coverage probabilities for different sample sizes ($n = 250, 500, 1000$). Additionally, we find that the performance of the procedure is not significantly affected by the choice of tuning parameters, including the number of folds used for cross-fitting and the bandwidth in nonparametric kernel estimates.

**6. Empirical Study**

In the empirical study, we apply the theoretical results to data from the Oregon Health Insurance Experiment (Finkelstein 2013), examining the potential outcomes of different types of
health insurance. To provide some background, the experiment was conducted by the state of Oregon between March and September 2008. A series of lottery draws were conducted to offer participants the option of enrolling in the Oregon Health Plan Standard, a Medicaid expansion program available to Oregon adult residents with limited income. Follow-up surveys were conducted in several waves to record various variables, including participants’ insurance plans and health status. Finkelstein et al. (2012) assessed the effects of insurance coverage after merging all types of insurance into a single treatment status and using a binary LATE model. We employ the GLATE model as a complimentary analysis to investigate the heterogeneity and substitution across different types of insurance.

To provide more details on the variables used, we examine three types of health-insurance-related outcomes: the number of primary care visits in the last six months, the number of days (out of the past 30) when poor health impaired regular activities, and the total out-of-pocket costs for medical care in the last six months. These outcomes are recorded by the Twelve Month Mail Survey conducted as part of the experiment. These outcome measures correspond to health care utilization, self-reported health, and financial strain, respectively. The covariates include household size and survey wave indicators, which must be controlled for to ensure lottery randomness. Additionally, we include lottery list variables and a pre-randomization measure of primary care visits.

For estimation, we follow the implementation procedure described in Section 4. We use $L_1$-penalized logistic regression to estimate the conditional instrument probabilities $\pi$, $L_1$-penalized multinomial logistic regression for the conditional treatment probabilities $P_{1,z}$, and lasso for the conditional expectations $Q_{1,z}$. To implement the lasso penalty, we set $\alpha = 1$ in the elastic net penalty. We employ 10-fold cross-validation to determine the regularization parameter for these penalized regressions. In the DML procedure, we use $L = 2$ folds for cross-fitting. The estimation outcomes are displayed in Table 2.

Besides the five insurance choice types presented in Table 1, we also include in the analysis the joint set of compliers, $\Sigma_{m,1}$, that consists of no- and $nm$-compliers. From the estimation results, we can see that half of the population are no-never-takers, who are never covered by any insurance plan. The compliers make up around one-fifth of the population. The size of $nm$-compliers is very small. Only 8.5% of the compliers would have otherwise chosen non-Medicaid insurance plans. The binary LATE estimates in Finkelstein et al. (2012) combine effects for these compliers with effects for compliers who would not otherwise have any insurance coverage. Since there are effectively no $nm$-compliers, our findings suggest that the experiment does not crowd out other insurance plan choices, which validates the interpretation of binary IV estimates as effects solely for individuals who would otherwise have no insurance coverage.

From the estimated LASFs, it is evident that there is a pattern of self-selection into insurance plans. By comparing the first and fourth rows in Table 2, we observe that, in a hypothetical scenario without any insurance coverage, no-compliers tend to have more primary care visits, nearly twice as many days with poor health, and similar out-of-pocket expenses compared to no-never-takers. Consequently, given the opportunity, no-compliers are more likely to choose the Medicaid program.

In the fifth row of Table 2, we notice that the standard errors of $\beta_{nn,1}$ are large, which aligns with our theoretical analysis considering the small size of the $nm$-compliers group. In this case, researchers can employ the robust inference procedures developed in the previous section to conduct any inference regarding the parameter $\beta_{nn,1}$.

By comparing the third and sixth rows in Table 2, we notice that always-takers, even when hypothetically covered by Medi-caid, tend to have more primary care visits, more days with poor health, and higher out-of-pocket expenses than compliers. One possible explanation for this observation lies in the nature of always-takers. As mentioned in footnote 23 by Finkelstein et al. (2012), always-takers’ ability to enroll without an offer must be attributed to very limited mechanisms, such as pregnant women staying on the plan after giving birth. As a result, always-takers could potentially be more susceptible to health and financial risks due to the unique circumstances that allow them to enroll in the plan.
Acknowledgments

This article is a revised version of the third chapter of my doctoral dissertation at UC San Diego. I am grateful to my advisors, Graham Elliott and Yixiao Sun, for their generous advice, guidance, and support. I appreciate the Editor Ivan Canay, an associate editor, and two referees whose valuable comments and suggestions have significantly improved this article. I also thank Wei-Lin Chen, Yu-Chang Chen, and Kaspar Wüthrich for insightful discussions.

Disclosure Statement

The author reports there are no competing interests to declare.

ORCID

Haitian Xie http://orcid.org/0000-0003-4284-081X

References

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Journal of Econometrics, 113, 231–263. [1056,1057]

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014), "Asymptotic Efficiency of Semiparametric Two-Step GMM," Review of Economic Studies, 81, 919–943. [1059]

Aliprantis, D., and Richter, F. G.-C. (2020), "Evidence of Neighborhood Effects from Moving to Opportunity: Lates of Neighborhood Quality," The Review of Economics and Statistics, 102, 633–647. [1053]

Andrews, D. W., Cheng, X., and Guggenberger, P. (2020), "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Journal of Econometrics, 218, 496–531. [1062]

Andrews, I., and Armstrong, T. B. (2017), "Unbiased Instrumental Variables Estimation Under Known First-Stage Sign," Quarterly Economics, 8, 479–503. [1062]

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," The Review of Economic Studies, 67, 499–527. [1054]

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," Journal of the American statistical Association, 90, 431–442. [1053,1054]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444–455. [1053]

Ao, W., Calonico, S., and Lee, Y.-Y. (2021), "Multivalued Treatments and Decomposition Analysis: An Application to the WIA Program," Journal of Business & Economic Statistics, 39, 358–371. [1054]

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993), Efficient and Adaptive Estimation for Semiparametric Models (Vol. 4), New York: Springer. [1054]

Cattaneo, M. D. (2010), "Efficient Semiparametric Estimation of Multivalued Treatment Effects Under Ignorability," Journal of Econometrics, 155, 138–154. [1054,1058]

Chen, X., and Christensen, T. M. (2015), "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," Journal of Econometrics, 188, 447–465. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao. [1061]

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric Efficiency in GMM Models with Auxiliary Data," The Annals of Statistics, 36, 808–843. [1054,1058]

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function is not Smooth," Econometrica, 71, 1591–1608. [1060]

Chen, X., and Santos, A. (2018), "Overidentification in Regular Models," Econometrica, 86, 1771–1817. [1058,1059]

Chen, Y.-C., and Xie, H. (2022), "Global Representation of the Conditional Late Model: A Separability Result," Oxford Bulletin of Economics and Statistics, 84, 789–798. [1055]

Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," The Econometrics Journal, 21, C1–C68. [1054,1060,1061]

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022), "Locally Robust Semiparametric Estimation," Econometrica, 90, 1501–1535. [1054,1060]

Farrell, M. H. (2015), "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," Journal of Econometrics, 189, 1–23. [1054]

Finkelstein, A. (2013), "Oregon Health Insurance Experiment Public Use Data," available at http://www.nber.org/oregon/data.html. [1062]

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012), "The Oregon Health Insurance Experiment: Evidence from the First Year," The Quarterly Journal of Economics, 127, 1057–1106. [1053,1054,1058,1063]

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012), "Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps," The Review of Economics and Statistics, 94, 153–171. [1054]

Friedman, J., Tibshirani, R., and Hastie, T. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," Journal of Statistical Software, 33, 1–22. [1061]

Frolíč, M. (2007), "Nonparametric iv Estimation of Local Average Treatment Effects with Covariates," Journal of Econometrics, 139, 35–75. [1054,1056,1057,1058,1059]

Galindo, C. (2021), "Essays on Treatment Effects from Multiple Unordered Choices," Ph. D. thesis, University of Maryland, College Park. [1053]

Goff, L. (2020), "A Vector Monotonicity Assumption for Multiple Instruments," arXiv preprint arXiv:2009.00553. [1054]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," Econometrica, 66, 315–331. [1054,1058,1059]

Hansen, B. E. (2008), "Uniform Convergence Rates for Kernel Estimation with Dependent Data," Econometric Theory, 24, 726–748. [1061]

Heckman, J. J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," The Review of Economic Studies, 65, 261–294. [1059]

Heckman, J. J., and Pinto, R. (2018), "Unordered Monotonicity," Econometrica, 86, 1–35. [1053,1054,1055,1057]

Heckman, J. J., and Vytlacil, E. (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica, 73, 669–738. [1054]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Econometrica, 71, 1161–1189. [1054]

Hong, H., and Nekipelov, D. (2010), "Semiparametric Efficiency in Nonlinear Models," Quantitative Economics, 1, 279–304. [1054,1058,1062]

Huber, M., Hsu, Y.-C., Lee, Y.-Y., and Lettry, L. (2020), "Direct and Indirect Effects of Continuous Treatments based on Generalized Propensity Score Weighting," Journal of Applied Econometrics, 35, 814–840. [1054]

Ichimura, H., and Newey, W. K. (2022), "The Influence Function of Semiparametric Estimators," Quantitative Economics, 13, 29–61. [1054]

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, 62, 467–475. [1054,1053,1057]

Imbens, G. W., and Manski, C. F. (2004), "Confidence Intervals for Partially Identified Parameters," Econometrica, 72, 1845–1857. [1062]

Kitagawa, T. (2015), "A Test for Instrument Validity," Econometrica, 83, 2043–2063. [1055]

Kline, P., and Walters, C. R. (2016), "Evaluating Public Programs with Close Substitutes: The Case of Head Start," The Quarterly Journal of Economics, 131, 1795–1848. [1053]

Lee, S., and Salanié, B. (2018), "Identifying Effects of Multivalued Treatments," Econometrica, 86, 1939–1963. [1054]
Lee, Y.-Y. (2018), “Efficient Propensity Score Regression Estimators of Multivalued Treatment Effects for the Treated,” *Journal of Econometrics*, 204, 207–222. [1054]

Masry, E. (1996), “Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates,” *Journal of Time Series Analysis*, 17, 571–599. [1061]

Mikusheva, A. (2007), “Uniform Inference in Autoregressive Models,” *Econometrica*, 75, 1411–1452. [1062]

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021), “The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables,” *American Economic Review*, 111, 3663–98. [1054]

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2023), “Policy Evaluation with Multiple Instrumental Variables,” *Journal of Econometrics*, forthcoming. [1054]

Mountjoy, J. (2022), “Community Colleges and Upward Mobility,” *American Economic Review*, 112, 2580–2630. [1053]

Navjeevan, M., Pinto, R., and Santos, A. (2023), “Identification and Estimation in a Class of Potential Outcomes Models,” arXiv preprint arXiv:2310.05311. [1058]

Newey, W. K. (1990), “Semiparametric Efficiency Bounds,” *Journal of Applied Econometrics*, 5, 99–135. [1054]

——— (1994), “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382. [1054]

Newey, W. K., and Stouli, S. (2021), “Heterogeneous Coefficients, Control Variables and Identification of Multiple Treatment Effects,” *Biometrika*, 109, 865–872. [1054]

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012), “Doubly Robust Instrumental Variable Regression,” *Statistica Sinica*, 22, 173–205. [1054]

Pinto, R. (2021), “Beyond Intention to Treat: Using the Incentives in Moving to Opportunity to Identify Neighborhood Effects,” UCLA working paper. [1053]

Simon, N., Friedman, J., Tibshirani, R., and Hastie, T. (2011), “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent,” *Journal of Statistical Software*, 39, 1–13. [1061]

Singh, R., and Sun, L. (2022), “Double Robustness for Complier Parameters and A Semiparametric Test for Complier Characteristics,” arXiv preprint arXiv:1909.05244. [1055]

Slóczynski, T., and Wooldridge, J. M. (2018), “A General Double Robustness Result for Estimating Average Treatment Effects,” *Econometric Theory*, 34, 112–133. [1055]

Stone, C. I. (1982), “Optimal Global Rates of Convergence for Nonparametric Regression,” *The Annals of Statistics*, 10, 1040–1053. [1061]

Sun, Z. (2023), “Instrument Validity for Heterogeneous Causal Effects,” *Journal of Econometrics* accepted. [1055]

Tan, Z. (2006), “Regression and Weighting Methods for Causal Inference Using Instrumental Variables,” *Journal of the American Statistical Association*, 101, 1607–1618. [1054]

Tay, J. K., Narasimhan, B., and Hastie, T. (2023), “Elastic Net Regularization Paths for All Generalized Linear Models,” *Journal of Statistical Software*, 106, 1–31. [1061]

Uysal, S. D. (2015), “Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling,” *Journal of Applied Econometrics*, 30, 763–786. [1055]

van der Vaart, A. W. (1998). *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [1059]

Vazquez-Bare, G. (2022), “Causal Spillover Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 118, 1911–1922. [1054]

Vytlacil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331–341. [1056]