Treatment Effect Estimation with Noisy Conditioning Variables

Kenichi Nagasawa[*]

April 28, 2024

Abstract

I develop a new identification strategy for treatment effects when noisy measurements of unobserved confounding factors are available. I use proxy variables to construct a random variable conditional on which treatment variables become exogenous. I show that, under appropriate conditions, there exists a one-to-one mapping between the distribution of unobserved confounding factors and the distribution of proxies, which in turn implies that holding constant the proxy distribution controls for the unobserved confounding factors. To ensure sufficient variation in the constructed control variable, I use an additional variable, termed excluded variable, which satisfies certain exclusion restrictions and relevance conditions. I establish asymptotic distributional results for semiparametric and flexible parametric estimators of causal parameters. I illustrate empirical relevance and usefulness of my results by estimating causal effects of grade retention on academic performance.

Keywords: Treatment effects, L^2-completeness, control functions, non-classical measurement errors.

[*] The University of Warwick, Department of Economics. Email: kenichi.nagasawa@warwick.ac.uk Address: Department of Economics, University of Warwick, Coventry, CV4 7AL, United Kingdom. This paper is based on a portion of my job market paper, which was circulated under the title "Identification and Estimation of Group-Level Partial Effects." I am grateful to my advisor Matias Cattaneo for advice and encouragement. I would like to thank Stéphane Bonhomme, Sebastian Calonico, Max Farrell, Yingjie Feng, Joachim Freyberger, Andreas Hagemann, Michael Jansson, Lutz Kilian, Xinwei Ma, Eric Renault, Rocío Titiunik, Gonzalo Vazquez-Bare, and seminar/conference participants at various institutions for helpful comments.

1 Introduction

In attempts to identify causal effects, researchers often exploit known treatment assignment rules. For instance, in some U.S. school districts, grade retention decisions are some function of standardized test scores. Economists have exploited such institutional feature to estimate causal effects of grade retention on various outcomes. In these settings, the success of treatment effect estimation relies on not only detailed knowledge of treatment assignment rules but also precise measures of characteristics used for treatment allocation. However, it is often difficult to have access to the exact variables used to determine treatment status due to data availability. Another such example arises in studying college admissions, where researchers rarely observe admission scores used by universities. In this paper, I develop a new identification strategy to address this type of empirical challenges. Specifically, I assume availability of proxy variables for unobserved variables that determine treatment status and devise a control function method to identify treatment effects. This new approach has a wide range of applications since coarse measurements of covariates are prevalent in practice (e.g., Gillen et al., 2019).

In its simplest form, the identification problem of interest is captured in the model

$$\begin{aligned} Y &= \beta_0 + \beta_1 D + \beta_2 X^* + \epsilon, \qquad \mathbb{E}[\epsilon(D, X^*)^\intercal] = 0, \\ &= \beta_0 + \beta_1 D + \beta_2 X + \varepsilon, \qquad \mathbb{E}[(X^* - X)(D, X^*, \epsilon)^\intercal] = 0, \end{aligned}$$

where $\varepsilon = \epsilon + \beta_2(X^* - X)$, D is the treatment of interest, X^* is the variable used for treatment assignment (unobserved to econometricians), and X is a proxy variable for X^*. If X^* were observed, one would estimate the equation $Y = \beta_0 + \beta_1 D + \beta_2 X^* + \epsilon$. However, the regression with X^* is infeasible, and the regression estimate using X in place of X^* is generally inconsistent for the treatment effect β_1. A textbook solution to this problem is to use a repeated measurement as an instrumental variable (IV) for X. Denoting the second measurement of X^* by Z, the two-stage least squares (2SLS) method is equivalent to estimating

$$Y = \beta_0 + \beta_1 D + \beta_2 \widehat{X} + \varepsilon, \qquad \mathbb{E}[\varepsilon(D, Z)^\intercal] = 0 \tag{1}$$

where $\widehat{X}$ is the linear projection of X on D, Z i.e., fitted value from the first-stage regression

$X = \gamma_0 + \gamma_1 D + \gamma_2 Z + \nu$. Here $\widehat{X}$ plays the role of a control function (Heckman and Robb, 1985); see Matzkin (2007, p.5356) for a definition of control functions.[1]

The main contribution of this paper is to extend the above control function method to non-separable models, including general nonparametric models. Specifically, I show that under appropriate assumptions, a vector V whose components take the form $\mathbb{E}[g(X)|D, Z]$ for some function g is a valid control function in the sense that the treatment variable becomes conditionally independent of unobserved confounding factors given V. This is a natural extension of (1) since my identification approach controls for general forms of fitted values relative to the 2SLS method. Accommodating non-separable models is especially important for treatment effect estimation because the linear model (1) imposes constant treatment effects, and as well-documented in the literature, ignoring treatment effect heterogeneity may produce misleading estimates.

The control function approach may be particularly appealing to applied researchers for its simplicity: the estimation procedure boils down to running regressions and computing sample averages. To facilitate implementation of the new control function method, I discuss semiparametric/flexible parametric modelling of the outcome equation, leveraging existing results in the control function literature. In addition, I develop a Lasso-based estimation procedure to flexibly choose regression specifications, building on Chernozhukov et al. (2022). With cross-validation, it is easy to choose Lasso tuning parameters, and I provide a closed-form variance estimator. I characterize a set of sufficient conditions for $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator as well as consistency of the variance estimator.

Relation to existing literature The main alternatives to my control function approach are the integral equation approach of Deaner (2018); Miao et al. (2018) and the operator diagonalization method of Hu and Schennach (2008). See Deaner (2022) for comparison of these two methods in the context of treatment effect estimation with measurement errors in covariates.

My identifying assumptions greatly overlap with those of the integral equation approach, yet there are some key differences in the identifying assumptions. In the integral equation approach, one imposes a high-level condition to ensure the existence of a solution to an integral equation,

[1] Here I take the perspective of control function rather than IV because the control function approach extends to non-linear settings naturally whereas IV counterparts face some difficulties (Blundell and Powell, 2003). For instance, nonparametric IV methods do not identify causal effects unless the unobserved heterogeneity is additively separable.

which I argue is difficult to verify except special cases. On the other hand, I instead impose a common support condition/overlap condition, which can be restrictive in practice but it can be empirically tested. I show in the supplemental appendix that with a similar high-level condition, my method yields an identification result analogous to the integral equation approach, and thus, this high-level condition is a key difference between the two approaches.

Hu and Schennach (2008) pioneered the use of completeness conditions in nonparametric measurement error models and their methods have been successfully applied beyond measurement error contexts (e.g., Arellano et al., 2017; Sasaki, 2015). Building on their idea, I also use a completeness condition to formalize the notion of a proxy variable. The conditional independence restrictions of the two approaches are different (Deaner, 2022), and thus, my result can complement the operator diagonalization method. Whereas Hu and Schennach (2008) identifies a larger class of parameters than my approach (e.g., the distribution of unobserved X^*), my control function method may be appealing to practitioners as estimation boils down to running regressions and computing sample averages.

This paper also contributes to the extensive literature on control functions (for reviews, see Blundell and Powell, 2003; Matzkin, 2007; Wooldridge, 2015) by proposing a novel construction of control functions. Many existing results construct a control variable from IV, whereas I use proxy variables. Since the identifying assumptions are quite different, the new result in this paper complements existing results by expanding the scope of applicability of control function methods. Many existing approaches use the invertibility of the first-stage equation to construct a control function, and this feature excludes the case of multi-dimensional unobserved heterogeneity with a scalar treatment variable. Provided that appropriate proxies are available, my method allows for multiple unobserved heterogeneity and does not require the invertibility of the first-stage equation. This aspect is practically relevant as imposing the scalar restriction on unobserved confounding factors may not be appealing in some settings. The requirement of first-stage invertibility also leads to a drawback that the existing methods have difficulty handling discrete endogenous variables (Wooldridge, 2015). My method is applicable to both discrete and continuous endogenous variables. One limitation I should note is that, similar to existing control function methods, my approach imposes possibly restrictive common support conditions, but I discuss how additional structures on the outcome equation can relax the common support condition.

I motivate my identification strategy as a generalization of 2SLS estimation using repeated measurements. While existing literature on non-linear models with measurement errors is extensive (for a review, see Schennach, 2020, and references therein), my approach is distinct from deconvolution methods. Also, Battistin and Chesher (2014) considered measurement error problems in covariates under small measurement error variance asymptotitcs.

Roadmap and notation In the next section, I describe the econometric model and present the identification results. Semiparametric and flexible parametric estimation methods are developed in Section 3, and I apply the results of this paper to estimating causal effects of grade retention on academic performance in Section 4. Section 5 concludes.

Let A, B be random variables. $f_{A,B}$ denotes a density of (A, B) with respect to some dominating measure, $f_{A|B}$ is the conditional density of A given B, $\text{supp}(A)$ denotes the support of A, and $\text{supp}(A|B = b)$ is the support of the conditional distribution of A given $B = b$. $\perp\!\!\!\perp$ denotes statistical independence and $A \perp\!\!\!\perp B|C$ means conditional independence between A and B given a random variable C.

2 Econometric model and identification results

$Y(d)$, $d \in \{0, 1\}$, denotes the potential outcome when treatment status is set equal to d (1 indicates treated and 0 otherwise), and D is the realized treatment status. I assume that the observed outcome Y satisfies $Y = DY(1) + (1 - D)Y(0)$. For simplicity, I focus on the binary treatment, but an extension to multi-valued treatment is straightforward. In settings considered below, the treatment allocation is a function of X^* and other idiosyncratic shocks. If X^* also affects the potential outcomes, then X^* is a confounding factor, and it is crucial to control for X^* in order to identify treatment effects. The main econometric challenge I focus on is that researchers may not observe X^* and instead they have a proxy variable X for X^*. For the identification strategy developed below, I rely on an additional variable that is similar to a repeated measurement in (1), denoted by Z. Since Z need not be a repeated measurement in my framework, I call Z an excluded variable, and precise assumptions on Z will be discussed below. For notational simplicity, I do not introduce covariates without measurement errors, but the identification analysis below goes through conditional on additional control variables.

To ground discussion on concrete terms, I use estimating causal effects of grade retention on future test scores as a running example (e.g., Fruehwirth et al., 2016; Schwerdt et al., 2017).

Example. The treatment of interest D is grade retention at a specific grade, say 1st grade, and the outcome of interest Y is a test score at a later stage e.g., one year later. By construction, retained students will be in a different grade from the control group, and we focus on comparing outcomes holding age constant. In this context, outcome measures are usually designed to allow for different-grade comparison. See the studies cited above for discussion on this point.

In the education economics literature, previous studies have exploited features of formal retention policies to identify causal effects. Most commonly, a grade retention policy stipulates that students whose standardized test scores fall below a cutoff will be retained at the current grade. That is, $D = \mathbb{1}\{\text{score} < \text{cutoff}\}$. For my identification strategy, I view the test score as a noisy measure of the underlying cognitive ability: $D = \mathbb{1}\{X^* + \eta < 0\}$ where X^* is (normalized) cognitive ability at the time of assessment and η is a "measurement error" in the standardized test score. Failure to control for X^* is likely to cause bias in treatment effect estimation because X^* affects Y, test performance one year later. Identification is challenging as researchers do not observe X^*. With access to standardized test scores used for grade retention decision, one can employ regression discontinuity designs and related strategies for identification. Here, I focus on the challenging case in which researchers do not have access to the test scores.

I assume that researchers observe noisy measures of cognitive ability: X denotes scores from tests that are administered independently of the standardized test used to determine grade retention. Z may represent parental investments in child's human capital (e.g., the number of books a student has at home) that affect test performance only indirectly through ability X^*. I further elaborate on requirements for Z below. Additional covariates such as observed student characteristics (e.g., household income) and pre-retention classroom variables (e.g., qualification of classroom teacher) may be used as additional controls, but I make them implicit for exposition. □

In this example, the treatment assignment is determined by a threshold rule. I emphasize that my identification strategy is not specific to this type of structure, and it is applicable to more general treatment assignment rules. For instance, suppose researchers have information on what variables (which may not be observed) are used for treatment allocation but might not have detailed knowledge on the functional form of the treatment assignment rule, and in this context, my identification strategy remains applicable as long as researchers have noisy measures of the key variables.

Parameters of interest include the average treatment effect (ATE) and the average treatment effect on the treated (ATT)

$$\mathtt{ATE} = \mathbb{E}[Y(1) - Y(0)], \qquad \mathtt{ATT} = \mathbb{E}[Y(1) - Y(0)|D = 1].$$

In the above example, the ATE represents the average effect of grade retention on future test scores, and the ATT focuses on the subpopulation of actually retained students. Given the nature of the intervention, grade retention may not be relevant for high-performing students, and thus, the ATT may be of direct interest in the context of grade retention.

As studied in the literature, other distributional features of the potential outcomes may be of interest e.g., the quantile and distributional structural functions. Also, one can consider conditional versions of the above objects e.g., average treatment effects conditional on student demographics. Using my approach, these objects are identifiable under similar assumptions that identify the ATE/ATT. As the identification arguments remain essentially identical, I focus on the average effects.

2.1 Identification result

For identification, I impose the following conditional independence restrictions.

Assumption 1. *For $d \in \{0, 1\}$, $Y(d) \perp\!\!\!\perp (D, Z)|X^*$.*

Assumption 2. *$X \perp\!\!\!\perp (D, Z)|X^*$.*

Assumption 1 is implied by $Y(d) \perp\!\!\!\perp D|X^*$ and $Y(d) \perp\!\!\!\perp Z|D, X^*$. The first part states that once conditional on X^*, the treatment variable becomes exogenous i.e., X^* represents confounding factors. The other condition $Y(d) \perp\!\!\!\perp Z|D, X^*$ states that Z satisfies an exclusion restriction in the sense that Z does not affect the outcome given the treatment and confounding factors. This exclusion restriction on Z is distinct from the standard IV exclusion restriction because the restriction needs to hold only conditional on X^*. Excluded variables Z are allowed to be correlated with the unobserved confounding factors, which is not the case for IVs.

Assumption 2 states that given the "correctly measured" variable, its noisy measurement is independent of other variables. This type of restriction is common in the measurement error literature (e.g., Assumption 2 in Hu and Schennach, 2008). As a visual aid to understanding Assumptions 1-2, in Figure 1, I present directed acyclical graphs that are compatible with the imposed conditional independence restrictions.

To assess the plausibility of the conditional independence restrictions in empirical settings, it is important to specify some elements of the treatment assignment mechanism. I illustrate this point using the grade retention example.

Example (continued)**.** The treatment assignment follows an explicit rule $D = \mathbb{1}\{X^* + \eta < 0\}$ where X^* is cognitive ability, and η is a measurement error in the standardized test score as noted above. In this context, $Y(d) \perp\!\!\!\perp D|X^*$ (part of Assumption 1) requires that the measurement error η be orthogonal to the test performance one year later, conditional on the cognitive ability at the time of the assessment for grade retention. Viewing η as an idiosyncratic shock, the assumption seems reasonable, especially because we condition on cognitive ability X^*. Similarly, the restriction $X \perp\!\!\!\perp (D, Z)|X^*$ (Assumption 2) holds under the orthogonality of measurement errors i.e., $X - X^*$ is independent of (η, Z) conditional on X^*.

The remaining part of Assumption 1 requires $Y(d) \perp\!\!\!\perp Z|D, X^*$, where Z represents parental investment in child's human capital. The rationale for this conditional independence restriction is that parental investment affects test performance only though its effects on child's ability. X^* is the cognitive ability at the time of test taking for grade retention decision, and Z denotes parental investment at the time period well before the test e.g., several months before. As in the model of Cunha et al. (2010), one can view parental investment Z as only an input to X^*, and once conditional on X^*, Z does not affect future test performance. □

Regarding Assumption 1, it is important that X^* captures all the sources of endogeneity. For example, if a grade retention decision is additionally based on teacher's assessment of student maturity, one may formulate the grade retention decision as

$$D = \mathbb{1}\{\text{academic test score} < \text{cutoff}\} \cup \{\text{maturity score} < \text{maturity cutoff}\}$$

where "maturity score" is teacher's evaluation of student's maturity level, which may not be observed by researchers. In this setting, X^* is a two-dimensional vector consisting of cognitive ability and personality traits (non-cognitive skills). If non-cognitive skills also affect the outcome of interest, it is important to expand X^* to include such characteristics.

Once researchers specify the elements of X^*, they need to identify proxy variables X that provide information about X^*. To formalize this idea, I follow Hu and Schennach (2008) and the subsequent literature using completeness conditions.

Assumption 3. *For any real-valued function b with $\mathbb{E}[b(X^*)^2] < \infty$, the distribution F_{XX^*} satisfies*

$$\mathbb{E}[b(X^*)|X] = 0 \quad F_X\text{-almost surely} \qquad \text{implies} \qquad b(X^*) = 0 \quad F_{X^*}\text{-almost surely.}$$

This condition is the L^2-completeness of the conditional distributions of X^* given X. Completeness conditions can be thought of as a nonparametric generalization of the IV rank condition in linear models (Newey and Powell, 2003). Since this is a rank condition, the dimension of X should be at least as large as that of X^* in general. In the literature, there are several known sufficient conditions for completeness (e.g., Andrews, 2017; D'Haultfoeuille, 2011; Hu et al., 2017). For instance, if researchers are willing to impose the measurement error structure such as $X = \chi(X^* + u)$ where χ is invertible and $X^* \perp\!\!\!\perp u$, then primitive sufficient conditions for completeness exist (see Lemma SA-4 in the appendix). In a panel data setting, Wilhelm (2015) discussed justifications for completeness assumptions using past observations as proxies. In the grade retention application, test scores are a strong indicator of underlying ability, and the rank condition seems reasonable.

The last substantive assumption is a version of overlap conditions.

Assumption 4. *Let $\mathcal{X} = \operatorname{supp}(X)$ and $\mathfrak{V} = \{f_{X|DZ}(x|D,Z) : x \in \mathcal{X}\}$. When the parameter of interest is ATE, $0 < \mathbb{P}[D = 1|\mathfrak{V}] < 1$ with probability one. When the parameter of interest is ATT, $\mathbb{P}[D = 1|\mathfrak{V}] < 1$ with probability one.*

As Theorem 1 below states, under Assumptions 1-3 and regularity conditions, the treatment assignment becomes exogenous conditional on $\mathfrak{V}$: $Y(d) \perp\!\!\!\perp D|\mathfrak{V}$. To identify treatment effects using this result, one needs to be able to vary D while holding constant $\mathfrak{V}$. Since $\mathfrak{V}$ is a function of D, this is possible only if the excluded variable Z can "undo" the effect of D on $\mathfrak{V}$, and this is the content of Assumption 4. This assumption differs from the standard overlap condition in that the propensity score is conditional on the stochastic process $\mathfrak{V} = \{f_{X|DZ}(x|D,Z) : x \in \mathcal{X}\}$. In the sequel, particularly in Section 3, I discuss how to estimate conditional expectations given $\mathfrak{V}$ under suitable assumptions on the data generating process.

Heuristically, Assumption 4 holds if for each z in $\mathcal{Z} = \text{supp}(Z)$, there is z' such that

$$f_{X|D,Z}(X|1,z) = f_{X|D,Z}(X|0,z') \tag{2}$$

with probability one.[2] As an example satisfying (2), consider the following model, which is consistent with the grade retention example. $D = \mathbb{1}\{X^* + \eta < 0\}$ where η is a random variable independent of (X^*, Z) following the standard logistic distribution, and the conditional distribution of X^* given Z belongs to an exponential family $f_{X^*|Z}(x^*|z) = \exp(\theta(z)x^* + A(z))h(x^*)$. Then, $f_{X^*|DZ}(x^*|d,z)$ equals

$$\mathbb{P}[D = d|X^* = x^*, Z = z]\frac{f_{X^*|Z}(x^*|z)}{\mathbb{P}[D = d|Z = z]} = \frac{\exp((1-d)x^*)}{1+\exp(x^*)}\frac{\exp(\theta(z)x^* + A(z))h(x^*)}{\mathbb{P}[D = d|Z = z]}$$

and

$$\frac{f_{X^*|DZ}(x^*|1,z)}{f_{X^*|DZ}(x^*|0,z')} \propto \exp\big((\theta(z) - \theta(z') - 1)x^*\big)$$

so in this model, the overlap condition holds if for each $z \in \mathcal{Z}$, there exists $z' \in \mathcal{Z}$ such that $\theta(z) = \theta(z') + 1$.[3] This example indicates that the overlap condition fails if the conditional distribution of X^* given $Z = z$ does not vary with z i.e., $\theta(z)$ is constant in z. In general, for the overlap condition to hold, the excluded variable should affect either D or X^* i.e., $\mathbb{P}[D = 1|X^*, Z]$ or $f_{X^*|Z}$ is a non-trivial function of Z.

For (2) to hold, it is essentially necessary for the conditional distribution to satisfy index restrictions: there exist some functions ψ and f such that

$$f_{X|DZ}(x|d,z) = f(x, \psi(d,z)) \tag{3}$$

[2] More precisely, for each $(d,z) \in \text{supp}(D,Z)$, there is $z' \in \text{supp}(Z|D = 1-d)$ such that $f_{X|DZ}(X|d,z) = f_{X|DZ}(X|1-d,z')$ holds with probability one.

[3] Under Assumption 2, (2) follows from $f_{X^*|D,Z}(X^*|1,z) = f_{X^*|D,Z}(X^*|0,z')$ almost surely.

for almost all $x, d, \tilde{z}$ and the dimension of $\psi(d,z)$ is at most $\dim(Z)$. This index restriction is essentially necessary because without such structure, we can find $z \in \mathcal{Z}$ for which there is no $z' \in \mathcal{Z}$ satisfying (2). In the above example, the index restriction also holds with $\psi(d,z) = 1 - d - \theta(z)$ since $\frac{\exp(A(z))}{\mathbb{P}[D=d|Z=z]}$ is some function of $\psi(d,z)$.

Another feature of the overlap condition is that it generally requires a large support of Z, a common restriction among control function approaches (e.g., Imbens and Newey, 2009, see their Assumption 2 and discussions following it). To see this requirement in the above example, suppose that $\theta(z) = z$. Then, if the support of Z is bounded, say $[a,b]$, then for $z \in [a, a+1)$, there is no z' in the support of Z satisfying $f_{X^*|D,Z}(\cdot|1,z) = f_{X^*|D,Z}(\cdot|0,z')$. The existing work on control function methods has recognized that the large support condition may be restrictive in practice (e.g., Chernozhukov et al., 2020; Florens et al., 2008). In the next subsection, I build on the insights from the literature to relax Assumption 4 by imposing additional structures on the outcome equation.

Before stating the identification result, I state a collection of regularity conditions.

Assumption 5. *$\mathbb{E}[|Y(d)|^2] < \infty$ for $d \in \text{supp}(D)$. With $\tilde{Z} = (D, Z)$, the joint distribution of $(X, X^*, \tilde{Z})$ is absolutely continuous with respect to the product measure $\lambda \times \lambda_* \times \lambda_{\tilde{Z}}$ where $\lambda, \lambda_*, \lambda_{\tilde{Z}}$ are σ-finite measures. $\mathbb{E}[|\frac{f_{XX^*}(X,X^*)}{f_X(X)f_{X^*}(X^*)}|^2] < \infty$ and $\int |\frac{f_{X^*|\tilde{Z}}(x^*|\tilde{Z})}{f_{X^*}(x^*)}|^2 f_{X^*}(x^*) d\lambda_*(x^*) < \infty$, $\int |f_{X|\tilde{Z}}(x|\tilde{Z})|^2 d\lambda(x) < \infty$ with probability one.*

Here is the main identification result of this paper. A formal proof is presented in the appendix.

Theorem 1. *Suppose Assumptions 1, 2, 3, and 5 hold. Then,*

$$Y(d) \perp\!\!\!\perp D | \mathfrak{V}$$

where $\mathfrak{V}$ is viewed as a random element in the space of λ-square integrable functions with the norm topology. In addition, if Assumption 4 holds, the ATE and ATT are identified by

$$\texttt{ATE} = \mathbb{E}[\mu(1, \mathfrak{V}) - \mu(0, \mathfrak{V})],$$

$$\texttt{ATT} = \mathbb{E}[\mu(1, \mathfrak{V}) - \mu(0, \mathfrak{V}) | D = 1],$$

where $\mu(D, \mathfrak{V}) = \mathbb{E}[Y|D, \mathfrak{V}]$.

The statement above focuses on average effects, but as standard in the literature, other parameters of interest such as the quantile structural function can be identified using the same argument. Also, the theorem focuses on the binary treatment, but one can extend it to multi-valued or continuously distributed D. For the extension, Assumptions 1-3 remain the same but Assumption 4 needs to be modified to the condition that the conditional support of $\mathfrak{V}$ given D equals the marginal support of $\mathfrak{V}$. This is essentially the same condition as Assumption 4, and in the supplemental appendix, I provide additional discussions on this support condition.

Conditional expectations given the stochastic process $\mathfrak{V}$ Here I discuss computation of conditional expectations given the stochastic process $\mathfrak{V}$, which is needed to use Theorem 1 for estimation. For further discussion of estimation issues, see Section 3.

In my setting, the stochastic process $\mathfrak{V}$ has a very specific structure that helps with computation. The sigma-field generated by $\mathfrak{V}$ is coarser than the sigma-filed generated by $(D, Z) =: \tilde{Z}$. Thus, intuitively, the effective dimension of $\mathfrak{V}$ is at most the dimension of $\tilde{Z}$ and potentially smaller. Moreover, the overlap condition (Assumption 4) essentially requires that the conditional distribution of X given $\tilde{Z}$ satisfies the index restriction (3). Then, a straightforward way to proceed is to estimate $\psi(D, Z)$ from data and use it as a control function in the outcome equation. That is, under Assumptions 1-5 and (3), $V_\psi = \psi(D, Z)$ is a valid control function in the sense of $Y(d) \perp\!\!\!\perp D|V_\psi$. This is an immediate corollary of Theorem 1 as $\sigma(V_\psi) \supseteq \sigma(\mathfrak{V})$, where $\sigma(\cdot)$ denotes the sigma-field generated by its argument.

Maintaining the index restriction, there is an alternative strategy to compute conditional expectations given $\mathfrak{V}$. Let $G(x) = (g_1(x), \ldots, g_k(x))^\intercal$ be a vector of F_X-integrable functions. Under (3),

$$\mathbb{E}[G(X)|\tilde{Z}] = \left(\int g_1(x) f(x, \psi(\tilde{Z})) d\lambda(x) \;\; \ldots \;\; \int g_k(x) f(x, \psi(\tilde{Z})) d\lambda(x) \right)^\intercal .$$

Since $\mathbb{E}[G(X)|\tilde{Z}]$ is a function of $V_\psi = \psi(\tilde{Z})$, $\sigma(\mathbb{E}[G(X)|\tilde{Z}]) \subseteq \sigma(V_\psi)$. On the other hand, if f is a smooth function of its second argument, with an appropriate choice of $G = (g_1, \ldots, g_k)^\intercal$ ($k \geq d_\psi$), the inverse function theorem implies that V_ψ can be expressed as a function of $\mathbb{E}[G(X)|\tilde{Z}]$ locally. By partitioning the probability space, we can turn this local result to a global one where conditioning on $\mathfrak{V}$ is equivalent to conditioning on $\mathbb{E}[G(X)|\tilde{Z}]$. Here is the formal assumption.

Assumption 6. *There exist $\psi : \mathrm{supp}(\tilde{Z}) \to \mathbb{R}^{d_\psi}$, $f : \mathrm{supp}(X, \psi(\tilde{Z})) \to [0, +\infty)$ such that $f_{X|\tilde{Z}}(x|\tilde{z}) = f(x, \psi(\tilde{z}))$ holds for almost all $x, \tilde{z}$, and the distribution of $V_\psi = \psi(D, Z)$ has a Lebesgue density. There exist square F_X-integrable functions $G(x) = (g_1(x), \ldots, g_k(x))^\intercal$ and a version of f in (3) such that $v \mapsto \int G(x)f(x, v)d\lambda(x)$ is continuously differentiable on $\mathrm{supp}(V_\psi)$ and $\frac{\partial}{\partial v^\intercal} \int G(x)f(x, v)d\lambda(x)$ is of column rank d_ψ for almost all v.*

With the structure in (3) (and Assumption 2), we can think of the conditional distribution of X given $(D, Z) = (d, z)$ as an element in the family of distributions $\{f(\cdot, \theta) : \theta \in \mathrm{supp}(V_\psi)\}$. The main restriction Assumption 6 imposes is that the parameter of this distribution is locally identifiable via the generalized method of moments using $\mathbb{E}_\theta[G(X)] = 0$ as the moment condition at almost all parameter values, where $\mathbb{E}_\theta$ is the expectation with respect to $f(\cdot, \theta)$. This restriction seems mild.

Researchers need to choose the functions G. One candidate is power functions i.e., $G(x) = (x, \ldots, x^k)^\intercal$, but any choice of G suffices provided that it satisfies the full rank condition in Assumption 6. Including more elements in G than necessary may lead to noisier estimates in a finite sample, but it can also guard against the potential violation of the full-rank condition. Thus, researchers may check robustness of estimation results by including additional functions to G.

Theorem 2. *Let $V = \mathbb{E}[G(X)|D, Z]$. Under Assumption 6, there exists a partition $\{A_j \in \sigma(\mathfrak{V}) : j \in \mathbb{N}\}$ of the underlying probability space such that for $\mathbb{P}_j(\cdot) = \mathbb{P}(\cdot \cap A_j)$,*

$$\mathbb{E}[Y|D, \mathfrak{V}] = \sum_{j \geq 1} \mathbb{E}_j[Y \mathbb{1}_{A_j}|D, V]\mathbb{1}_{A_j} \qquad \textit{almost surely}$$

where conditional expectations using $\mathbb{P}_j$ is defined in the same way as to the case using a probability measure.

This theorem implies that one can estimate the conditional expectation $\mathbb{E}[Y|D, \mathfrak{V}]$ by (nonparametrically) regressing Y on (D, V) where V is estimated in the first stage. The presence of a partition $\{A_j : j \in \mathbb{N}\}$ can be handled by using a local basis approximation such as piecewise polynomials. The partition is not unique, and as long as a user-chosen partition is finer than a theoretical partition, the local approximation enables estimation. I discuss further implementation details in Section 3.

2.2 Semiparametric/parametric outcome equation

As previously mentioned, the overlap condition (Assumption 4) can be restrictive in applications. Building on the control function literature, I discuss additional structures researchers can impose on the outcome equation to relax the support condition. I should emphasize that while the modelling technique discussed below is borrowed from the literature, my contribution is the novel construction of the control function using a proxy variable.

In the sequel, I write V to denote either $\psi(D, Z)$ or $\mathbb{E}[G(X)|D, Z]$ as both are valid control functions under appropriate assumptions. Deviating from the grade retention example, I also consider multi-valued/continuous D.

Random coefficient model I consider a version of the random coefficient model

$$Y = \varepsilon_1 + \varepsilon_2 D, \qquad (\varepsilon_1, \varepsilon_2) \perp\!\!\!\perp D | X^* \tag{4}$$

with $D \in \mathbb{R}$ having more than two support points i.e., multi-valued treatment or continuously distributed variable. This model allows for heterogeneous treatment effects as the treatment variable interacts with the unobserved heterogeneity that may be correlated with X^*. The causal parameter of interest I consider is the average of the random slope on the treatment variable:

$$\theta_0 = \mathbb{E}[\varepsilon_2].$$

This parameter represents the average marginal effect of the treatment D. Under Assumptions 1-3 and 5, the model (4) yields

$$\mathbb{E}[Y|D, V] = \mu_1(V) + \mu_2(V)D$$

where μ_1, μ_2 are unknown functions. Newey and Stouli (2021) showed that if the conditional variance of D given $V = v$ is positive for almost all v, then θ_0 is identified under regularity conditions. This new identifying condition is substantially weaker than the common support condition: while the common support condition demands that $\text{supp}(D|V = v)$ equals $\text{supp}(D)$ almost surely, the new condition only requires $\text{supp}(D|V = v)$ has more than one point almost surely. Therefore, with the random coefficient specification (4), an excluded variable Z may have discrete variation and one can still achieve the identification of the causal parameter θ_0.[4]

With a binary treatment D, the model (4) places no restrictions on the outcome determination process, and one needs to impose additional structures to relax the overlap condition. One possibility is to impose that $\mu_1(v), \mu_2(v)$ are real analytic and the conditional support of V given D contains an open set. This identification argument was used by Arellano and Bonhomme (2017). A nice feature of this approach is that the assumptions encompass the cases where $\mu_1(v), \mu_2(v)$ are polynomial functions of v, a widely used specification in empirical studies.

Flexible parametric model Let $q(d,v)$ be a vector of some transformations of (d,v) e.g., $q(d,v) = (1,d,v,dv)^\intercal$. If desired, higher-order polynomial transformations may be included in $q(d,v)$. I postulate that the outcome equation satisfy

$$\mathbb{E}[Y|D,V] = \Lambda\big(q(D,V)^\intercal \gamma_0\big) \tag{5}$$

where Λ is a known, strictly monotonic link function and γ_0 is the parameter to be estimated. This model encompasses a parametric version of the random coefficient model considered above. Suppose

$$Y = \varepsilon_1 + \varepsilon_2 D, \qquad \mathbb{E}[\varepsilon_1|D,V] = p(V)^\intercal \gamma_1, \quad \mathbb{E}[\varepsilon_2|D,V] = p(V)^\intercal \gamma_2$$

where $p(v)$ is a vector of transformations of v and the conditional independence $\varepsilon_l \perp\!\!\!\perp D|V$, $l = 1,2$ follows under Assumptions 1-3 and 5. Relative to the above semiparametric model, the additional restriction here is $\mathbb{E}[\varepsilon_l|V] = p(V)^\intercal \gamma_l$ for $l = 1,2$. This random coefficient model fits into (5), where Λ equals the identity function, $q(D,V) = (p(V)^\intercal, Dp(V)^\intercal)^\intercal$, and $\gamma_0 = (\gamma_1^\intercal, \gamma_2^\intercal)^\intercal$.

To provide another example of (5), consider the binary outcome model

$$Y = \mathbb{1}\{\gamma_0 D \geq \varepsilon\}, \qquad \varepsilon \perp\!\!\!\perp (D,Z)|X^*.$$

Suppose that $F_{\varepsilon|X^*}(e|x^*) = F(e + \delta^\intercal x^*)$ for some function F and coefficient vector $\delta \in \mathbb{R}^{\dim(X^*)}$ and that $f_{X^*|DZ}(x^*|d,z) = f(x^* - \iota(d,z))$ for some fixed functions f, ι. Also, $X = X^* + U_x$ with

[4] Let $\nu(D,Z) = \mathbb{E}[G(X)|D,Z] = V$. For $v \in \mathrm{supp}(V)$, if there exist two distinct points $(d_1,z_1),(d_2,z_2) \in \mathrm{supp}(D,Z)$ such that $v = \nu(d_1,z_1) = \nu(d_2,z_2)$, then the conditional variance of D given $V = v$ is positive.

$U_x \perp\!\!\!\perp (X^*, D, Z)$ and $\mathbb{E}[U] = 0$. Then, $V = \mathbb{E}[X|D, Z] = \iota(D, Z)$ and (5) holds with

$$\Lambda(\cdot) = \int F(\cdot + \delta^{\mathsf{T}} u) f(u) du, \qquad q(D, V) = (D, V^{\mathsf{T}})^{\mathsf{T}}.$$

With $\dim(X^*) = 1$, if F and f are standard normal cdf and density respectively, then $\Lambda(y) = F(y/\sqrt{1 + \delta^2})$, implying a probit model.

Under (5), identification of treatment effects holds if the parameter γ_0 is identified. In turn, the coefficients γ_0 are identified if the matrix $\mathbb{E}[q(D, V)q(D, V)^{\mathsf{T}}]$ is non-singular. This full-column rank condition is weaker than the support invariance condition, and Chernozhukov et al. (2020) and Newey and Stouli (2021) provided sufficient conditions for the non-singularity of $\mathbb{E}[q(D, V)q(D, V)^{\mathsf{T}}]$. In the binary outcome example, the non-singularity of $\mathbb{E}[q(D, V)q(D, V)^{\mathsf{T}}]$ holds if for each $d \in \text{supp}(D)$, there exist z_1, z_2 in $\text{supp}(Z|D = d)$ such that $\iota(d, z_1) \neq \iota(d, z_2)$.

3 Estimation

I propose a semiparametric estimator for average treatment effects (ATEs) and flexible parametric estimators for average causal effects. I provide a set of sufficient conditions for asymptotic normality of the estimators and propose consistent variance estimators.

3.1 Semiparametric estimation

Theorem 2 indicates that instead of conditioning on $\mathfrak{V}$, it suffices to condition on a finite-dimensional vector $V = \mathbb{E}[G(X)|D, Z]$ with some choice of $G(x) = (g_1(x), \dots, g_k(x))^{\mathsf{T}}$ and a partition based on (D, Z). While there are generally many valid choices of G, researchers do not know a priori what choice of G is appropriate. In this situation, a practical procedure is to initially include many candidate functions and use a model selection procedure to choose relevant ones. To develop a formal theory, I focus on a Lasso-based estimation method, building on the recent developments using Neyman-orthogonal moments. The treatment here mostly follows Chernozhukov et al. (2022) (henceforth CNS). In my setting, there is an additional complication due to the need to estimate the control function V, a generated regressor problem, and I extend the theory of CNS to handle the complication. For omitted details of the Neyman-orthogonal moment theory, I refer interested readers to CNS. Below, I focus on ATE as the parameter of interest. Although the discussion in this section focuses on this specific estimand, the estimation method below can be extended to other causal parameters and multi-valued treatment settings.

Let $\mu_0(d, v) = \mathbb{E}[Y|D = d, V = v]$ and $V = \nu_0(D, Z)$ with $\nu_0(D, Z) = \mathbb{E}[G(X)|D, Z]$ where $G(x)$ potentially contains redundant elements, which may be discarded via a model selection procedure. For the theory below, the choice of G is fixed in the asymptotics. Writing θ_0 for ATE, we have

$$\theta_0 = \mathbb{E}[\mu_0(1, V) - \mu_0(0, V)] \equiv \mathbb{E}[m(\Xi, \mu_0, \nu_0)]$$

where $\Xi = (Y, D, Z^\intercal, X^\intercal)^\intercal$ is a data observation and $m(\xi, \mu, \nu) = \mu(1, \nu(d, z)) - \mu(0, \nu(d, z))$ is the moment function. Having characterized the moment function m, a simple approach to estimate θ_0 is to first estimate μ_0, ν_0 and form a sample analogue of $\mathbb{E}[m(\Xi, \mu_0, \nu_0)]$. Yet, such plug-in approach is known to suffer from bias arising from noises in the first-stage estimates of (μ_0, ν_0), especially when one uses Lasso or other machine learning methods to select regressors. Neyman-orthogonal estimation equations can be used to address this issue.

To describe the Neyman-orthogonal moment, let

$$\alpha_0(\Xi) = \frac{D}{\mathbb{P}[D = 1|V]} - \frac{1 - D}{1 - \mathbb{P}[D = 1|V]}.$$

CNS refers to this function as a Riesz representer, and it plays an important role in the theory developed by CNS. Using the pathwise derivative calculation of Hahn and Ridder (2013) and Newey (1994) with the assumptions imposed below, one can show that

$$\begin{aligned}\psi(\Xi, \mu, \nu, \alpha) = m(\Xi, \mu, \nu) &+ \alpha(D, \nu(D, Z))[Y - \mu(D, \nu(D, Z))] \\ &+ \left[\frac{\partial}{\partial V^\intercal} m(\Xi, \mu, \nu) - \alpha(D, \nu(D, Z)) \frac{\partial}{\partial V^\intercal} \mu(D, \nu(D, Z))\right] [G(X) - \nu(D, Z)]\end{aligned}$$

is an orthogonal score function for the parameter θ_0.[5] ψ being an orthogonal score means that (i)

[5] The form of the score function crucially depends on the auxiliary result $\mathbb{E}[Y|D, Z] = \mathbb{E}[Y|D, \mathfrak{V}]$ almost surely (c.f., Hahn et al., 2022). See Lemma A.3 in the supplemental appendix.

$\mathbb{E}[\psi(\Xi, \mu_0, \nu_0, \alpha_0)] = \theta_0$ and (ii) for a path $\{\mu_t, \nu_t, \alpha_t : t \in [0, \epsilon), \epsilon > 0\}$,

$$\frac{\partial}{\partial t}\mathbb{E}[\psi(\Xi, \mu_t, \nu_t, \alpha_t)]\big|_{t=0} = 0$$

holds. This second property suggests that the score function is insensitive to the first-stage estimation errors in (μ_0, ν_0, α_0). Then, one can form an estimator of θ_0 by first estimating (μ_0, ν_0, α_0) and plugging them into ψ to form a sample analogue of $\mathbb{E}[\psi(\Xi, \mu_0, \nu_0, \alpha_0)]$.

Following CNS, I use cross-fitting. Let $\{I_l\}_{l=1}^{L}$ be a L-fold partition of the sample. The number of partition L is fixed (a common choice is $L = 5$, or 10), and the size of each partition should be similar. Estimation involves multiple steps. In the first step, I estimate the control function $V = \nu_0(D, Z)$. Any nonparametric/machine learning method may be used to construct estimates of V, and I write $\widehat{V}_{ll'} = \widehat{\nu}_{ll'}(D, Z)$ for the estimate of V using observations not in $I_l, I_{l'}$. Here, all the elements in $\mathbb{E}[G(X)|D, Z]$ are estimated, some of which may be redundant and discarded later.

In the second step, I estimate (μ_0, α_0) using partitioning-based least squares estimation (e.g., Cattaneo et al., 2020) and Lasso. Let $p(v) = (p_1(v), \ldots, p_K(v))^\intercal$ be a vector of locally supported approximating functions (e.g., piecewise polynomials) whose dimension $K = K_n$ depends on the sample size. Define

$$\widehat{\Omega}_l = \frac{1}{n - n_l}\sum_{l' \neq l}\sum_{i \in I_{l'}}\begin{bmatrix} p(\widehat{V}_{i,ll'})p(\widehat{V}_{i,ll'})^\intercal & p(\widehat{V}_{i,ll'})p(\widehat{V}_{i,ll'})^\intercal D_i \\ p(\widehat{V}_{i,ll'})p(\widehat{V}_{i,ll'})^\intercal D_i & p(\widehat{V}_{i,ll'})p(\widehat{V}_{i,ll'})^\intercal D_i \end{bmatrix}$$

$$\widehat{M}_l^{\mu} = \frac{1}{n - n_l}\sum_{l' \neq l}\sum_{i \in I_{l'}}\begin{bmatrix} p(\widehat{V}_{i,ll'})Y_i \\ p(\widehat{V}_{i,ll'})D_i Y_i \end{bmatrix}, \qquad \widehat{M}_l^{\alpha} = \frac{1}{n - n_l}\sum_{l' \neq l}\sum_{i \in I_{l'}}\begin{bmatrix} \mathbf{0} \\ p(\widehat{V}_{i,ll'}) \end{bmatrix}$$

where n_l is the size of I_l. Then, the Lasso estimators of (μ_0, α_0) are given by $\widehat{\mu}_l(d, v) = (p(v)^\intercal, p(v)^\intercal d)\widehat{\rho}_l$, $\widehat{\alpha}_l(d, v) = (p(v)^\intercal, p(v)^\intercal d)\widehat{\delta}_l$ where

$$\widehat{\rho}_l = \arg\min_{\rho}\left\{\rho^\intercal\widehat{\Omega}_l\rho - 2\rho^\intercal\widehat{M}_l^{\mu} + 2\kappa_n\|\rho\|_1\right\}, \qquad \widehat{\delta}_l = \arg\min_{\delta}\left\{\delta^\intercal\widehat{\Omega}_l\delta - 2\delta^\intercal\widehat{M}_l^{\alpha} + 2\kappa_n\|\delta\|_1\right\},$$

κ_n is a sequence of penalty terms that shrinks to zero, and $\|\cdot\|_1$ is the ℓ^1-norm on $\mathbb{R}^{2K}$. Finally, an estimator of the causal parameter is formed by

$$\widehat{\theta}_n = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\psi(\Xi_i, \widehat{\mu}_l, \widehat{\nu}_l, \widehat{\alpha}_l)$$

where $\widehat{\nu}_l$ is an estimate of ν_0 using observations not in I_l. Also, one can estimate the asymptotic variance of $\widehat{\theta}_n$ by

$$\widehat{\Psi}_n = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left[\psi(\Xi_i, \widehat{\mu}_l, \widehat{\nu}_l, \widehat{\alpha}_l) - \widehat{\theta}_n\right]^2.$$

The use of a locally supported approximation basis is partly motivated by Theorem 2, which states that the conditional expectation of interest can be expressed as the regression function of (D, V) with some partition defined by (D, Z). For concreteness, take piecewise polynomials as the approximating basis $p(\cdot)$. Then, a researcher uses rectangles to partition the sample space $\{(D_i, Z_i) : i = 1, \ldots, n\}$ and on each rectangle, one fits a regression of Y on polynomials of $\widehat{V}$ and their interactions with D. For details on piecewise polynomials and other local approximation bases, see Cattaneo et al. (2020). In my setting, unlike the standard procedure, while partitioning is based on $\{(D_i, Z_i) : i = 1, \ldots, n\}$, regressors are D and polynomial functions of $\widehat{V}$. Yet, the theoretical analysis remains unchanged because V is a function of (D, Z) and as each rectangle shrinks at an appropriate rate, the basis can approximate any smooth functions.[6]

To present a formal result on the asymptotic distribution of $\widehat{\theta}_n$, I define additional notations. Write $\|\cdot\|_j$ for the ℓ^j-norm $j = 1, 2$ on Euclidean spaces. Let $q(d, v) = (p(v)^\intercal \; dp(v)^\intercal)^\intercal$ be the vector of approximating functions. For $u = (u_1, \ldots, u_l) \in \mathbb{Z}^l_{\geq 0}$, write $|u| = \sum_{\ell=1}^{k} u_\ell$ and $\partial^u f(x) = \partial^{|u|} f(x)/\partial^{u_1} x_1 \ldots \partial^{u_l} x_l$. Below $C > 1$ denotes a positive constant independent of the sample size and represents a different number at different places.

Assumption 7. *Observations $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ are a random sample. μ_0 is twice continuously differentiable in v and α_0 is Lipschitz continuous in v. μ_0, α_0, and derivatives of μ_0 are bounded. Let $\varepsilon = Y - \mu_0(D, V)$. With probability one, $\mathbb{E}[\varepsilon^2 | D, Z] \leq C < \infty$. Also, $\mathbb{E}[\varepsilon^4] < \infty$.*

This assumption imposes regularity conditions on the underlying data generating process, which

[6] The use of a local approximation basis is convenient because choosing a partition determines the tuning parameter K_n. However, one can also use a partition whose element does not shrink to a point and use a global approximation basis (e.g., polynomials) on each element of the partition. This approach is theoretically valid because an appropriate partition in Theorem 2 is fixed (independent of the sample size). The assumptions below are stated in a way that both approaches are accommodated.

are mild and standard in the literature. The next assumption imposes that the first-stage estimate of the control function is consistent and converges at a certain rate in L^2 norm.

Assumption 8. *The first-stage estimator $\widehat{\nu}_n$ is uniformly bounded and satisfies $\int \|\widehat{\nu}_n(t,z) - \nu_0(t,z)\|_2^2 dF_{DZ}(t,z) = O_{\mathbb{P}}(\tilde{\omega}_n^2)$ where $\tilde{\omega}_n = o(1)$ is the convergence rate of $\widehat{\nu}_n$.*

Define

$$\chi_{a,n} = \max_{1 \le l \le K_n} \max_{|u|=a} \sup_{v \in \mathrm{supp}(V)} |\partial^u p_l(v)| \qquad a = 0, 1, 2,$$

where $\partial^0 p(u) = p(u)$. Then, let

$$\omega_n = \chi_{1,n} \tilde{\omega}_n.$$

This rate plays an important role in the assumption below. With specific $p(\cdot)$, bounds on $\chi_{a,n}$ are available in the literature. For instance, with polynomial splines, $\chi_{a,n} = K_n^{1/2+a}$ (Newey, 1997).

The next assumption needs some additional notation: given a set of indices $J \subset \{1, \ldots, L\}$ and $v \in \mathbb{R}^L$, let $\#|J|$ be the cardinality of J, v_J be the $\#|J| \times 1$ subvector of v whose elements are $\{v_l : l \in J\}$, and v_{J^c} be the subvector of v that consists of elements not in v_J.

Assumption 9. *The basis function $p(\cdot)$ is twice continuously differentiable. The eigenvalues of $\mathbb{E}[p(V)p(V)^\intercal]$ and $\Omega = \mathbb{E}[q(D,V)q(D,V)^\intercal]$ are bounded above uniformly in K. Ω and its sample analogue $\widehat{\Omega}$ satisfy that for any $s = O(\omega_n^{-2})$, with probability approaching one,*

$$\min_{\substack{J \subset \{1,\ldots,2K\} \\ \#|J| \le s}} \min_{\|\rho_{J^c}\|_1 \le 3\|\rho_J\|_1} \frac{\rho^\intercal \widehat{\Omega} \rho}{\rho_J^\intercal \rho_J} \ge c, \qquad \min_{\substack{J \subset \{1,\ldots,2K\} \\ \#|J| \le s}} \min_{\|\rho_{J^c}\|_1 \le 3\|\rho_J\|_1} \frac{\rho^\intercal \Omega \rho}{\rho_J^\intercal \rho_J} \ge c$$

where $c > 0$ is independent of K.

This condition is a sparse eigenvalue condition commonly used in the Lasso literature (see Assumption 3 of CNS).

Assumption 10. *Let $\bar{\rho}, \bar{\delta}$ be least squares coefficients of projecting μ_0, α_0 onto $q(D,V)$, respectively. i.e., $\Omega\bar{\rho} = \mathbb{E}[q(D,V)Y]$ and $\Omega\bar{\delta} = \mathbb{E}[q(D,V)\alpha_0(D,V)]$. There exists $\zeta > 1/2$ such that for each positive integer s, there exist $\widetilde{\rho}, \widetilde{\delta} \in \mathbb{R}^{2K_n}$ satisfying that the number of non-zero elements in $\widetilde{\rho}, \widetilde{\delta}$ are bounded by s and*

$$\|\bar{\rho} - \widetilde{\rho}\|_2 + \|\bar{\delta} - \widetilde{\delta}\|_2 \le Cs^{-\zeta}.$$

Let $\bar{\mu}(d,v) = q(d,v)^\intercal \bar{\rho}$ and $\bar{\alpha}(d,v) = q(d,v)^\intercal \bar{\delta}$. For $t \in \{0,1\}$,

$$\mathbb{E}[|\mu_0(t,V) - \bar{\mu}(t,V)|^2] + \mathbb{E}[|\mu_0(D,V) - \bar{\mu}(D,V)|^2] + \mathbb{E}[|\alpha_0(D,V) - \bar{\alpha}(D,V)|^2] = o(n^{-1/2}),$$
$$\mathbb{E}[\|\partial\{\mu_0(t,V) - \bar{\mu}(t,V)\}/\partial V\|_2^2] + \mathbb{E}[\|\partial\{\mu_0(D,V) - \bar{\mu}(D,V)\}/\partial V\|_2^2] = o(1)$$
$$\sup_{v\in\text{supp}(V)} |\mu_0(t,v) - \bar{\mu}(t,v)| + \sup_{v\in\text{supp}(V)} |\alpha_0(t,v) - \bar{\alpha}(t,v)| = o(1).$$

Assumption 10 imposes that μ_0, α_0 admit sparse approximations. I follow Bradic et al. (2022) in the formulation of this condition (their Assumptions 3 and 8). There are two aspects that differ from CNS: (L^2-) approximation of the derivative of μ_0 and uniform approximations of μ_0 and α_0. These approximation properties hold for many of standard choices of approximating functions.

Theorem 3. *Assumptions 1-10 hold and $\chi_{\ell,n} \leq \chi_{\ell+1,n}$ for $\ell = 0,1$. With $\tilde{\kappa}_n = \kappa_n/\omega_n$, suppose $\max_{1\leq l\leq L} n/n_l = O(1)$, $K_n/n + \tilde{\kappa}_n^{-1} + \chi_{0,n}\max\{\omega_n^{(2\zeta-1)/(2\zeta+1)}, [n^{1/4}\log K_n]^{-1}\} + \chi_{1,n}\omega_n^{2\zeta/(2\zeta+1)}\tilde{\kappa}_n + \chi_{2,n}\tilde{\omega}_n = o(1)$, and $\omega_n^{2\zeta/(2\zeta+1)}\tilde{\kappa}_n + (\chi_{0,n}\chi_{2,n})^{1/2}\tilde{\omega}_n = o(n^{-1/4})$. Then,*

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \rightsquigarrow \text{Normal}(0, \Psi_0)$$

where $\Psi_0 = \text{Var}[\psi(\Xi, \mu_0, \nu_0, \alpha_0)]$ and $\rightsquigarrow$ denotes convergence in distribution. Also, the variance estimator $\widehat{\Psi}_n$ converges in probability to Ψ_0.

The theorem states that the estimator $\widehat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically normal under the rate restrictions on the first-stage estimator, the Lasso penalty term, and the sup norm on the $\partial^u p(v)$ with $|u| \leq 2$. To illustrate, suppose we use polynomial splines as approximating functions and $\tilde{\omega}_n = n^{-\tau_1}$, $K_n = n^{\tau_2}$ for some $\tau_1, \tau_2 > 0$. Then, the hypothesis of Theorem 3 implies $2\tau_1 - 3\tau_2 > \frac{1}{2} + \frac{1}{4\zeta}$. The restriction is somewhat stringent on the rate at which the number of approximation terms can grow (i.e., K_n) because both the first and second stages are nonparametric in V. It might be possible to relax rate restrictions and to obtain a better finite-sample distributional approximation by carefully analyzing the higher-order terms of the estimator (e.g., in the spirit of Cattaneo et al., 2019). I leave such analysis for future work.

3.2 Flexible parametric approach

In this section, I consider a flexible parametric estimation procedure based on the model (5) in Section 2.2. Recall that (5) posits

$$\mathbb{E}[Y|D,V] = \Lambda(q(D,V)^{\mathsf{T}}\gamma_0)$$

where $\Lambda(\cdot), q(\cdot)$ are specified by researchers and γ_0 is to be estimated. For the control function V, I posit the model

$$V = \mathbb{E}[G(X)|D,Z] = Q(D,Z)\delta_0,$$

where $Q : \text{supp}(D,Z) \to \mathbb{R}^{d_v \times d_q}$ is a user-chosen matrix-valued transformation of (D,Z) and $\delta_0 \in \mathbb{R}^{d_q}$ is the parameter to be estimated. As a baseline, one may use $Q(D,Z) = I \otimes q_2(D,Z)$ with I being the identity matrix, $q_2(D,Z) = (1, D^{\mathsf{T}}, Z^{\mathsf{T}})$ and $\otimes$ denoting the Kronecker product. Researchers can include higher-order polynomial terms to enhance flexibility.

For implementation, first estimate δ_0 by least squares and form $\widehat{V}_i = Q(D_i,Z_i)\widehat{\delta}_n$. Then, estimate γ_0 in $\mathbb{E}[Y|D,V] = \Lambda(q(D,V)^{\mathsf{T}}\gamma_0)$ by (non-linear) regression of Y on $q(D,\widehat{V})$. Finally, the estimator for $\theta_0(d) = \mathbb{E}[Y(d)]$ is formed by

$$\widehat{\theta}_n(d) = \frac{1}{n}\sum_{i=1}^{n} \Lambda\big(q(d,\widehat{V}_i)^{\mathsf{T}}\widehat{\gamma}_n\big).$$

One can form an estimate of the ATE by $\widehat{\theta}_n(1) - \widehat{\theta}_n(0)$. For binary treatment, by forming the sum over treated individuals i.e., $\sum_{i=1}^{n} D_i\Lambda(q(d,\widehat{V}_i)^{\mathsf{T}}\widehat{\gamma}_n)/\sum_{i=1}^{n} D_i$, one can construct an estimator for ATT as well.

For inference, it is useful to have a closed-form variance estimator. Let $\dot{\Lambda}$ be the first derivative of Λ,

$$\widehat{q}_i = q(D_i,\widehat{V}_i), \quad \widehat{\varepsilon}_i = Y_i - \Lambda(\widehat{q}_i^{\mathsf{T}}\widehat{\gamma}_n), \quad Q_i = Q(D_i,Z_i), \quad \widehat{\eta}_i = G(X_i) - Q_i\widehat{\delta}_n,$$

$$\widehat{\Gamma}_1 = \frac{1}{n}\sum_{i=1}^{n} Q_i^{\mathsf{T}}Q_i, \quad \widehat{\Gamma}_2 = \frac{1}{n}\sum_{i=1}^{n} \big|\dot{\Lambda}(\widehat{q}_i^{\mathsf{T}}\widehat{\gamma}_n)\big|^2\widehat{q}_i\widehat{q}_i^{\mathsf{T}}, \quad \widehat{\Gamma}_3 = -\frac{1}{n}\sum_{i=1}^{n} |\dot{\Lambda}(\widehat{q}_i^{\mathsf{T}}\widehat{\gamma}_n)|^2\widehat{q}_i\widehat{\gamma}_n^{\mathsf{T}}\partial q(D_i,\widehat{V}_i)Q_i,$$

$$\widehat{c}_1(d) = \frac{1}{n}\sum_{i=1}^{n} \dot{\Lambda}(q(d,\widehat{V}_i)^{\mathsf{T}}\widehat{\gamma}_n)q(d,\widehat{V}_i)^{\mathsf{T}}, \qquad \widehat{c}_2(d) = \frac{1}{n}\sum_{i=1}^{n} \dot{\Lambda}(q(d,\widehat{V}_i)^{\mathsf{T}}\widehat{\gamma}_n)\widehat{\gamma}_n^{\mathsf{T}}\partial q(d,\widehat{V}_i)Q_i,$$

and ∂q be the derivative of q with respect to V. Then,

$$\widehat{\Psi}_n(d) = \frac{1}{n}\sum_{i=1}^{n}\left[\Lambda(q(d,\widehat{V}_i)^\intercal\widehat{\gamma}_n) - \widehat{\theta}_n(d) + \widehat{c}_1(d)\widehat{\Gamma}_2^{-1}\widehat{q}_i\dot{\Lambda}(\widehat{q}_i^\intercal\widehat{\gamma}_n)\widehat{\varepsilon}_i + \{\widehat{c}_1(d)\widehat{\Gamma}_2^{-1}\widehat{\Gamma}_3 + \widehat{c}_2(d)\}\widehat{\Gamma}_1^{-1}Q_i^\intercal\widehat{\eta}_i\right]^2$$

is an estimator for the asymptotic variance of $\sqrt{n}(\widehat{\theta}(d) - \theta(d))$. Note that the variance estimator does not require additional nuisance parameter estimation.

Since the asymptotic distributional theory for $\widehat{\theta}_n(d)$ is well-established (e.g., Newey and McFadden, 1994), I relegate the discussion of the asymptotic theory to the appendix. Under the assumptions stated there, the estimator $\widehat{\theta}_n(d)$ is asymptotically normal and the variance estimator is consistent.

4 Empirical application

I apply the results developed in this paper to studying causal effects of grade retention on short-term academic performance. I use W to denote the vector of observed control variables such as student and school characteristics.

4.1 Data description

I use data from Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K), a longitudinal study that followed a sample of students from kindergarten to the fifth grade. The study sample consists of children enrolled in kindergarten during the 1998-1999 school year, and the initial sampling was designed to obtain a nationally representative sample in the United States. The first year of the longitudinal study (referred to as Year 1 in the sequel) corresponded to the 1998-1999 school year. In Year 2, most students were enrolled in the first grade while some students (approximately 3%) repeated the kindergarten grade. In ECLS-K, students were assessed on their knowledge in reading, mathematics, and general science in the fall and spring of Year 1 and the spring of Year 2. Identical assessment batteries were used regardless of student's retention status. I use the measures in Year 1 as proxy variables for the unobserved cognitive skill and the measures in Year 2 as the outcomes of interest.

For the choice of regression controls, I build on Fruehwirth et al. (2016), who used ECLS-K data to study causal effects of grade retention over a longer time horizon. Covariates include student/family characteristics and classroom/school characteristics. In addition to proxy variables for cognitive skills, the dataset contains (noisy) measures of student's personality traits. I also observe measures of parental investment in child's human capital. The variables used in analysis are listed in Tables 1 and 2 along with summary statistics. Exploiting the availability of multiple measurements of student ability, I use some proxy variables measured at the fall of Year 1 as part of Z, for which I discuss the plausibility of the identifying assumptions below. The full sample of ECLS-K contains 21,409 students and after dropping units with missing observations, the final sample consists of 6,277 students, out of whom 205 students repeated the kindergarten grade. Relative to students who proceeded to the first grade, retained students have lower test scores on average across subjects and time of test administration. The treated group has the lower averages of Socioeconomic Status measure and child's age, has a higher percentage of male students, and has more books at home while the averages of other variables are mostly comparable between the treated and control groups.

In the ECLS-K sample, students attended schools across various U.S. school districts, and I do not have good information on how grade retention decisions were made. Then, I conduct the empirical analysis under the assumption that the grade retention decision was a function of student cognitive ability, student maturity, and other idiosyncratic shocks. That is, letting $X^* = (X_1^*, X_2^*)^\intercal$ where X_1^* and X_2^* denote cognitive ability and maturity level evaluated by school teachers, I assume that $D = h(X^*, W, \eta)$ where h is some fixed function, η is an idiosyncratic shock, and W is the vector of observed controls. One concern about this assumption is that schools are likely to vary in their retention policies. In an attempt to address this issue, I include in W school characteristics as well as variables that describe school's formal retention policies e.g., whether schools retain students based on maturity or parent's request.

Plausibility of the identifying assumptions have been discussed in Section 2.1. I provide additional comments regarding the assumptions related to measurement errors. In the present context, X^* denotes student's ability at the end of Year 1 when a grade retention decision is made, Z contains proxy variables measured in the fall of Year 1, and X come from proxy variables measured in the spring of Year 1. Then, $Y(d) \perp\!\!\!\perp Z | X^*$ (part of Assumption 1) seems reasonable because conditional on ability at the end of Year 1, test scores in early part of Year 1 are less likely to have predictive power for test scores in Year 2. For Assumption 2, X consist of test scores measured in the spring of Year 1, and a key assumption is that these test scores did not directly affect the grade retention decision, conditional on the underlying ability X^*. According to Deaner (2023, p.35), test scores in the ECLS-K study were not shared with the students, parents, nor teachers, and thus, it is plausible that these measures did not directly affect the grade retention decision. Also, since I use some proxy variables in Z, Assumption 2 requires that the measurement errors in X and those in the test scores included in Z be independent. This restriction may be plausible given that test scores in X were measured at a different time from those in Z. If X and Z were to include test scores measured on the same day, the measurement errors could be correlated e.g., students were feeling unwell on the day of test taking and performed systematically worse than if they had not been sick.

As the parameter of interest, I focus on the average treatment effects on the treated (ATTs). Because grade retention is most relevant for academically struggling students, ATTs are more natural objects to consider than the average treatment effects, which are based on the distribution of all students. For comparison, I also estimate ATTs using proxy variables as additional controls. Under my econometric framework, using proxy variables as controls does not fully address the endogeneity issue because measurement errors induce bias in the estimates.

4.2 Model specifications

I use the modelling approach discussed in Sections 2.2 and 3.2. For the outcome equation, I use the specification (5) with the identity link function.

For the specification of the control function V, I use two approaches. The first one is the regression specification as discussed in Section 3.2. Specifically, let $X = (X_1, X_2)^\intercal$ where X_1 is a proxy for cognitive ability and X_2 is a noisy measure of student maturity, and I specify

$$V = \big(\mathbb{E}[X_1|D, Z, W], \mathbb{E}[X_2|D, Z, W]\big)^\intercal \tag{6}$$

i.e., $G(x) = (x_1, x_2)^\intercal$ in the notation of the previous sections. Also, I model the regression functions by linear specifications.

For the second approach, I use the framework of Spady and Stouli (2020). I briefly describe their framework and refer interested readers to the original paper for details. For random variables $U_1 \in \mathbb{R}, U_2 \in \mathbb{R}^k$, their modelling strategy starts from the observation that if the conditional distribution of U_1 given U_2 is continuous,

$$\Phi^{-1}\big(F_{U_1|U_2}(U_1|U_2)\big) =_d \text{Normal}(0,1)$$

where Φ is the cdf of standard normal. The model primitive is the function $g(u_1, u_2) \equiv \Phi^{-1}(F_{U_1|U_2}(u_1|u_2))$, and they propose to use the specification

$$g(u_1, u_2) = S_1(u_1) \otimes S_2(u_2)\delta_0$$

where S_1, S_2 are vectors of transformations whose first element is unity and δ_0 is the parameter to be estimated. If $S_1(u_1) = 1$ and $S_2(u_2) = (1, u_2^\intercal)^\intercal$, then g function corresponds to the case where $F_{U_1|U_2}(u_1|u_2) = \Phi(\frac{u_1 - u_2^\intercal\beta}{\sigma})$ i.e., Gaussian case. Including additional terms in S_i's allows for flexible modelling of the conditional distribution. For my application, letting $\tilde{Z} = (D, Z, W)$, I model $F_{X_1|X_2\tilde{Z}}$ and $F_{X_2|\tilde{Z}}$ using Spady and Stouli (2020)'s approach, where I impose restrictions on the parameters so that each of $F_{X_1|X_2\tilde{Z}}(x_1|x_2, \tilde{z})$ and $F_{X_2|\tilde{Z}}(x_2|\tilde{z})$ satisfies the single-index restriction as in (3). Then, as discussed above, I can use the two-dimensional random vector $\psi(\tilde{Z})$ as a valid control function.

I use the reading score as X_1 and the self-control measure as X_2, both of which were measured in the spring of Year 1. For Z, I use measures of parental investment in child's human capital as well as math, approach-to-learning, and interpersonal scores, all of which were measured in the fall of Year 1. These proxies exhibit large average gaps across the treated and control groups, which heuristically motivated the choice. More importantly, the choices of these proxy variables are dictated by the restrictions that only scores measured in the spring of Year 1 be used as X, that Z only contain scores from the fall of Year 1, and that both cognitive and behavioral measures be included in X and Z.

In the next section, I present ATT estimates along with some robustness checks on the above specifications as well as testing some of the model implications. Given the availability of repeated measures, I test the conditional independence restriction involving measurement errors. Namely, Assumption 2 imposes that measurement errors in test scores of the spring of Year 1 be statistically independent of the treatment status conditional on X^*. The proof of Theorem 1 indicates that under Assumptions 1-3, proxy variables should be orthogonal to the treatment indicator D once conditional on the control function. Since I have four measurements taken in the spring of Year 1 that were not used in estimation, I conduct an empirical test of the above model implication by regressing the unused test scores on the treatment variable and the control function. The use of left-out measurements is motivated by the idea that using X both in the control function estimation and in the testing might lead to spurious results in a moderate sample size. I should note that the empirical test only considers the restriction on averages, and failure to reject the null hypothesis of conditional mean independence does not guarantee that Assumption 2 holds.

4.3 Estimation results

In Table 3, I present the results of the linear regressions of unused proxies on the treatment, the control function, and other controls. If my identifying assumptions hold, we should see that the coefficient estimate be not statistically different from zero. For both specifications of the control functions, the coefficients on the treatment dummy, except for one, are not distinguishable from zero. For the index specification, the result is consistent with the assumption that the measurement errors are orthogonal to the treatment status given X^*, and the same is true for the regression specification except for one test score. I proceed with both specifications, and as seen below, the estimation results are mostly comparable across the specifications.

Figure 2 contains the histogram of estimated propensity score. I estimate the propensity score by logistic regression using the control function with the index specification and other covariates. From the histogram, the propensity score is bounded away from 1 but it is not bounded away from 0. Because I focus on ATTs, Assumption 4 seems to hold.

As a further specification check, I conduct the conditional Kolmogorov-Smirnov test of Andrews (1997) to test whether the modelling approach of Spady and Stouli (2020) is appropriate with the proxy variables in the ECLS-K sample. Specifically, I test whether the parametric model of the conditional distribution of X given (D, Z, W) fits well with the data. Using 1,000 bootstrap iterations, the p-value of the test statistic is 0.88; I do not reject the parametric specification of the conditional distribution.

In Table 4, I present the estimates of ATTs and the associated 95% confindence intervals. The outcomes are standardized so the ATT estimates can be interpreted as changes in standard deviation. On the first column, the estimates based on OLS using proxy variables as additional controls i.e., this specification controls for all the past test scores. Due to measurement errors, I expect that the OLS estimates suffer from bias. For the remaining columns, the estimates based on my control function method with different specifications are presented. The columns "(1)" "(2)" "(4)" present estimates using the index specification of V, where the "(1)" is based on regression with a linear specification, "(2)" is based on regression including square and interaction terms, and "(4)" contain estimates based on inverse probability weighting (IPW). The columns "(3)" "(5)" present estimates using the mean specification (6), where "(3)" is based on regression with square and interaction terms and "(5)" is for IPW estimates.

For the general knowledge test, the OLS estimate suggests a statistically significant negative effect of grade retention (0.09 standard deviations) while the control function estimates suggest that the effect is indistinguishable from 0. For the math score, the causal effect estimates based on OLS indicate that the grade retention effect is substantial i.e., 0.46 standard deviations. In contrast, the control function methods suggest the causal effect is much smaller than the OLS estimate i.e., point estimates around 0.3 standard deviations, although the IPW estimates are closer to the OLS estimate. For reading, the differences between the OLS and control function estimates are smaller than for the other subjects, and yet, the control function estimates still indicate that the effects are smaller than what the OLS estimate might suggest.

The existing studies in the economics of education literature suggest that short-term causal effects of grade retention are often null or slightly positive. For example, Jacob and Lefgren (2009) observed that observational studies using non-experimental designs found negative associations between grade retention and academic achievements whereas their causal estimates of short-run effects were not distinguishable from zero. In light of the findings in the existing empirical literature, the above estimation result might suggest that the proposed control function method corrects for the bias due to unobserved ability, at least partially if not completely. Thus, this empirical application indicates the importance of properly dealing with measurement errors in covariates.

5 Conclusion

I developed a new identification strategy for causal effects such as the average treatment effects by exploiting proxy variables for unobserved confounding factors. Using this new result, I provided a simple, rgression-based method to estimate causal effects. Specifically, I developed a Lasso-based method to flexibly choose regression specifications for my control function method. As illustrated through an empirical application, measurement errors in covariates can have non-negligible impact on causal estimates.

6 Bibliography

Andrews, D. W. K. (1997): "A Conditional Kolmogorov Test," *Econometrica*, 65, 1097–1128.

——— (2017): "Examples of L^2-Complete and Boundedly-Complete Distributions," *Journal of Econometrics*, 199, 213–220.

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734.

Arellano, M. and S. Bonhomme (2017): "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85, 1–28.

Battistin, E. and A. Chesher (2014): "Treatment Effect Estimation with Covariate Measurement Error," *Journal of Econometrics*, 178, 707–715.

Blundell, R. W. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnsovsky, Cambridge University Press, vol. 2, chap. 8, 321–357.

Bradic, J., V. Chernozhukov, W. K. Newey, and Y. Zhu (2022): "Minimax Semiparametric Learning with Approximate Sparsity," Working paper.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020): "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, 48, 1718–1741.

Cattaneo, M. D., M. Jansson, and X. Ma (2019): "Two-Step Estimation and Inference with Possibly Many Included Covariates," *Review of Economic Studies*, 86, 1095–1122.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2020): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," *Quantitative Economics*, 11.

Chernozhukov, V., W. Newey, and R. Singh (2022): "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90, 967–1027.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Deaner, B. (2018): "Proxy Controls and Panel Data," Working Paper.

——— (2022): "Controlling for Latent Confounding with Triple Proxies," Working Paper.

——— (2023): "Proxy Controls and Panel Data," Working Paper.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206.

Fruehwirth, J. C., S. Navarro, and Y. Takahashi (2016): "How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects," *Journal of Labor Economics*, 34, 979–1021.

Gillen, B., E. Snowberg, and L. Yariv (2019): "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study," *Journal of Political Economy*, 127, 1826–1863.

Hahn, J., Z. Liao, G. Ridder, and R. Shi (2022): "The Influence Function of Semiparametric Two-Step Estimators with Estimated Control Variables," Working paper.

Hahn, J. and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica*, 81, 351–340.

Heckman, J. J. and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. S. Singer, Cambridge University Press, Econometric Society Monographs, 156–246.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y., S. M. Schennach, and J.-L. Shiu (2017): "Injectivity of a Class of Integral Operators with Compactly Supported Kernels," *Journal of Econometrics*, 200, 48–58.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations models without Additivity," *Econometrica*, 77, 1481–1512.

Jacob, B. A. and L. Lefgren (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1, 33–58.

Matzkin, R. L. (2007): "Nonparametric identification," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5307 – 5368.

Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018): "Identifying Causal Effects with Proxy Variables of an Unmeasured Confounder," *Biometrika*, 105, 987–993.

Newey, W. and S. Stouli (2021): "Control Variables, Discrete Instruments, and Identification of Structural Functions," *Journal of Economietrics*, 222, 73–88.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111 – 2245.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Sasaki, Y. (2015): "Heterogeneity and Selection in Dynamic Panel Data," *Journal of Econometrics*, 188, 236–249.

Schennach, S. M. (2020): "Mismeasured and Unobserved Variables," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 487–565.

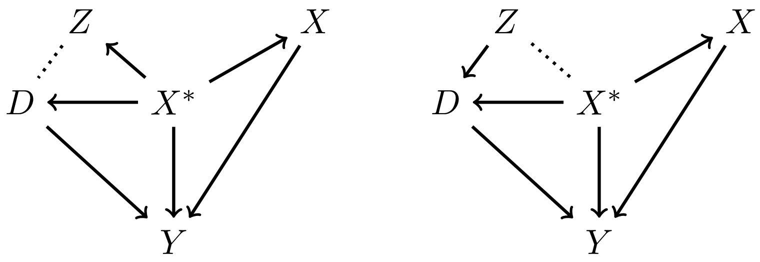
Schwerdt, G., M. R. West, and M. A. Winters (2017): "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida," *Journal of Public Economics*, 152, 154–169.

Spady, R. H. and S. Stouli (2020): "Gaussian Transforms Modeling and the Estimation of Distributional Regression Functions," Working paper.

Wilhelm, D. (2015): "Identification and Estimation of Nonparametric Panel Data Regression with Measurement Error," CEMMAP Working Paper.

Wooldridge, J. M. (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445.

Figure 1: Independence Assumptions via DAG



Notes. Arrows represent causal effects and dotted lines mean that two variables may have a causal relationship with unspecified direction of effects.

Table 1: Summary statistics

	All: N=6277		**Control: N=6072**		**Treated: N=205**	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
Outcome						
General score Y2	0.217	0.870	0.237	0.860	−0.396	0.948
Math score Y2	0.153	0.985	0.191	0.968	−0.983	0.793
Reading score Y2	0.124	0.971	0.166	0.954	−1.124	0.556
Covariates						
SES measure	0.131	0.759	0.139	0.757	−0.105	0.798
Body mass index	16.236	2.110	16.242	2.117	16.062	1.896
Teacher experience	14.454	9.027	14.470	9.019	13.995	9.273
Class size	20.457	4.974	20.493	4.957	19.390	5.340
Male	0.501	0.500	0.496	0.500	0.654	0.477
White	0.678	0.467	0.678	0.467	0.673	0.470
Black	0.118	0.323	0.118	0.323	0.132	0.339
Hispanic	0.108	0.311	0.108	0.311	0.107	0.310
Age	5.850	0.690	5.857	0.696	5.634	0.418
Kindergarten full-time	0.570	0.495	0.569	0.495	0.612	0.487
# of siblings	1.406	1.081	1.400	1.075	1.585	1.252
TV rule at home	0.893	0.309	0.893	0.310	0.902	0.297
Absence of father	0.164	0.370	0.163	0.369	0.195	0.397
Absence of mother	0.014	0.118	0.014	0.117	0.020	0.139
% minority in school:						
1-5%	0.206	0.404	0.207	0.405	0.171	0.377
5-10%	0.157	0.364	0.159	0.366	0.112	0.316
10-25%	0.101	0.301	0.100	0.300	0.122	0.328
>25%	0.124	0.330	0.124	0.330	0.137	0.344
Public school	0.780	0.414	0.783	0.412	0.698	0.460
Title 1 Funding	0.620	0.485	0.621	0.485	0.595	0.492
Crime is a concern	0.428	0.562	0.426	0.562	0.512	0.548
Students bring weapon	0.164	0.370	0.166	0.372	0.102	0.304
Teacher/student attacked	0.372	0.483	0.372	0.483	0.371	0.484
Security measure	0.055	0.228	0.055	0.227	0.068	0.253
Parent involvment	2.998	0.900	3.000	0.897	2.912	0.976
Teacher has MA degree	0.350	0.477	0.351	0.477	0.307	0.463
Class behavior rating	1.564	0.787	1.560	0.786	1.668	0.815
% minority in classroom:						
1-5%	0.083	0.276	0.083	0.276	0.083	0.276
5-10%	0.130	0.337	0.130	0.337	0.137	0.344
10-25%	0.195	0.396	0.195	0.396	0.200	0.401
>25%	0.390	0.488	0.390	0.488	0.385	0.488
Retention policy:						
For immatuiry	0.763	0.425	0.763	0.425	0.766	0.425
Needs parent approval	0.452	0.498	0.452	0.498	0.429	0.496
For parent request	0.761	0.426	0.762	0.426	0.746	0.436

Notes. Outcome test scores were standardized to have zero mean and unit variance before the sample construction process.

Table 2: Summary statistics: continued

	All: N=6277		**Control: N=6072**		**Treated: N=205**	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
Proxy variables						
General score Y1S	0.187	0.941	0.207	0.935	−0.418	0.914
Math score Y1S	0.176	0.973	0.204	0.968	−0.677	0.667
Reading score Y1S	0.087	0.988	0.114	0.989	−0.705	0.537
App.to learning score Y1S	0.127	0.954	0.157	0.939	−0.757	0.975
Self-control score Y1S	0.105	0.963	0.119	0.960	−0.316	0.982
Interpersonal score Y1S	0.115	0.968	0.131	0.966	−0.337	0.929
General score Y1F	0.168	0.970	0.190	0.965	−0.482	0.863
Math score Y1F	0.159	0.989	0.186	0.988	−0.647	0.602
Reading score Y1F	0.077	0.974	0.099	0.979	−0.583	0.493
App.to learning score Y1F	0.142	0.962	0.168	0.951	−0.621	0.981
Self-control score Y1F	0.094	0.962	0.106	0.959	−0.279	0.980
Interpersonal score Y1F	0.095	0.966	0.107	0.965	−0.273	0.925
Excluded variables						
# of books at home	83.429	60.269	83.724	60.246	74.683	60.431
# of CDs at home	16.552	17.995	16.614	17.963	14.707	18.887
How often parents:						
Read to child	3.304	0.748	3.306	0.746	3.244	0.810
Tell stories	2.743	0.913	2.741	0.909	2.800	1.026
Sing songs together	3.124	0.909	3.127	0.907	3.034	0.977
Help doing art	2.660	0.846	2.661	0.845	2.615	0.870
Involve child in chores	3.300	0.846	3.297	0.846	3.390	0.825
Play games together	2.790	0.797	2.789	0.794	2.815	0.872
Talk about nature	2.248	0.855	2.245	0.853	2.337	0.891
Build together	2.361	0.902	2.356	0.898	2.483	1.003
Play sport together	2.683	0.888	2.682	0.888	2.727	0.888
Look at picture book	3.354	0.779	3.357	0.778	3.259	0.802
Read on their own	2.977	0.900	2.986	0.898	2.732	0.940
Watch Sesame Street	0.574	0.495	0.574	0.495	0.576	0.495

Notes. All the test scores (proxy variables) were standardized to have zero mean and unit variance before the sample construction process. Y1S stands for “Year 1 Spring” and Y1F stands for “Year 1 Fall”.

Table 3: Testing the conditional independence restriction on measurement errors

	General score	Math score	App.to.learn score	Interpersonal score
Index Specf.				
Coef. Est.	0.092	0.123	−0.091	0.014
Std. err.	0.098	0.139	0.123	0.124
95% CI	[−0.113 0.263]	[−0.189 0.368]	[−0.362 0.129]	[−0.244 0.242]
Reg. Specf.				
Coef. Est.	−0.005	0.067	−0.168	0.018
Std. err.	0.051	0.041	0.062	0.051
95% CI	[−0.106 0.096]	[−0.016 0.149]	[−0.287 −0.040]	[−0.080 0.120]

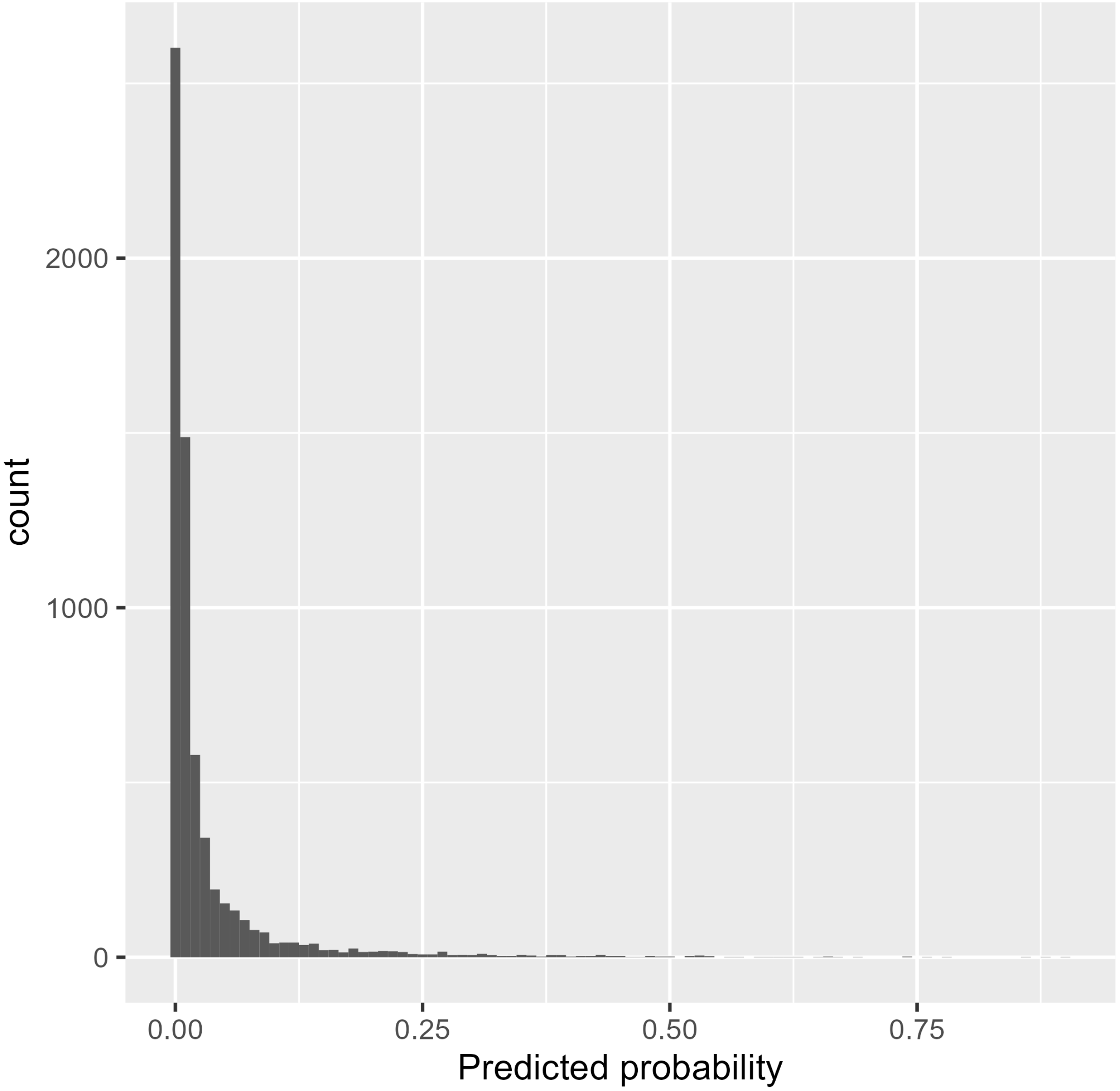
Notes. The estimated model is $\tilde{X} = \alpha + \beta D + \gamma^{\mathsf{T}} V + \delta^{\mathsf{T}} W + \epsilon$ where $\tilde{X}$ equals either general subject score or interpersonal skills score. The top panel presents the regression coefficient estimates using V with the index specification. The bottom panel shows the results for V specified by (6). The standard errors and quantiles for the confidence intervals were computed using 1,000 bootstrap iterations.

Table 4: ATT estimates

	OLS (ME bias)	CF (1)	CF (2)	CF (3)	CF (4)	CF (5)
Geneal						
Est.	−0.088	0.009	0.020	−0.019	0.019	−0.025
95% CI	[−0.161 −0.014]	[−0.190 0.179]	[−0.186 0.207]	[−0.136 0.089]	[−0.198 0.229]	[−0.155 0.106]
Math						
Est.	−0.461	−0.290	−0.339	−0.326	−0.394	−0.406
95% CI	[−0.536 −0.386]	[−0.592 −0.080]	[−0.628 −0.091]	[−0.431 −0.223]	[−0.589 −0.183]	[−0.517 −0.291]
Reading						
Est.	−0.572	−0.446	−0.489	−0.486	−0.560	−0.573
95% CI	[−0.635 −0.510]	[−0.696 −0.262]	[−0.719 −0.323]	[−0.564 −0.401]	[−0.715 −0.395]	[−0.651 −0.481]

Notes. The column “OLS (ME bias)” contains estimates based on regression using proxy variables as controls. The columns “CF (1)” “CF (2)” “CF (4)” present estimates using the index specification of V, where the “CF (1)” is based on regression with a linear specification, “CF (2)” is based on regression including square and interaction terms, and “CF (4)” contain estimates based on inverse probability weighting (IPW). The columns “CF (3)” “CF (5)” present estimates using the mean specification (6), where “CF (3)” is based on regression with square and interaction terms and “CF (5)” is for IPW estimates. The quantiles for the confidence intervals were computed using 1,000 bootstrap iterations.

Figure 2: Histogram of Estimated Propensity Score



Notes. The propensity score is estimated by logistic regression using the control variables and the estimated V variable, specified by the index restriction.