From One LATE to Another: Machine Learning and the External Validity of IV Estimates*

Alexander Kwon† Kyungtae Lee‡

November 13, 2026

Abstract

Instrumental variable (IV) estimates identify local average treatment effects (LATEs) for compliers, whose composition can differ across populations. We study when a LATE estimated in a sample population can predict the LATE in a target population. We decompose the prediction error of the benchmark model, interacted two-stage least squares (I2SLS), into four terms and study which terms a more flexible model, Double/Debiased Machine Learning (DML), can reduce. Among them, we show that terms related to nonlinearity in the instrument propensity and outcome equations are significantly reduced by DML, but two terms, one related to the nonlinearity of the Conditional Average Treatment Effect and the other related to compliance weights, are not. We find that this decomposition is important because the overall prediction error can be small even when the individual terms are large, as they can cancel each other out due to differences in their signs. We apply these results in two Monte Carlo simulations, including one calibrated to data, and find supporting evidence.

JEL Codes: C26, C55, C52

Keywords: External Validity, Instrumental Variables, Local Average Treatment Effect, Machine Learning, Double/Debiased ML

*Authors are listed in alphabetical order and contributed equally to this work. We thank participants at the AFES 2026 conference.

†Graham School of Business, York College of Pennsylvania. Email: akwon@ycp.edu

‡Ph.D. Program in Economics, Graduate Center, CUNY. Email: kyungtae.lee79@gc.cuny.edu

1 Introduction

Empirical evidence from instrumental variable (IV) studies often guides policy decisions beyond the original study setting. Yet IV estimates identify local average treatment effects (LATEs) that may not generalize when the composition of compliers differs across populations (Imbens and Angrist, 1994; Angrist and Pischke, 2009). In this paper, we study when a LATE estimated in one population can be used to predict the LATE in another. We provide conditions under which the target LATE is identified using information from the sample population and target covariates, decompose the prediction error of a Two-Stage Least Squares benchmark, and present which components flexible machine-learning methods can reduce and which can remain.

The LATE-to-LATE prediction problem arises naturally in policy. Consider a government that has estimated the return to schooling using college proximity as an instrument in one region (Card, 1995) and wants to predict the return in a new region before expanding geographic access to college there. The relevant compliers are people whose enrollment responds to college proximity. Regional differences in observed covariate distributions can make direct transfer of average treatment effects inaccurate (Allcott, 2015; Hotz et al., 2005). For example, when treatment effects are heterogeneous, differences in the complier distribution can make the prediction of the target LATE using the estimated sample LATE inaccurate, with direct consequences for policy evaluation and targeting.[1]

In this paper, we use a linear interacted IV specification as a benchmark, which we call interacted two-stage least squares (I2SLS), the term used by Zhao et al. (2025).[2] We then decompose its target-LATE prediction error into four components. We then introduce Double/Debiased Machine Learning (DML), which flexibly estimates the nuisance functions and uses Neyman-orthogonal moments (Chernozhukov et al., 2018), to study which of these

[1] Section 1.1 illustrates the problem of predicting a target site LATE from sample-site evidence using data on solid fuel use and cooking time in six developing countries.

[2] This approach builds on the cross-site prediction framework of Allcott (2015) and the IV implementation in Kwon and Lee (2025); up to differences in covariate centering, Zhao et al. (2025) provide a theoretical analysis of the interacted 2SLS specification within a single population.

components can be reduced by more flexible estimation and which can remain even when the nuisance functions are estimated more accurately.

We make three contributions. First, we derive a bias decomposition (Proposition 2) showing that the prediction error of an estimated linear conditional complier effect consists of four additive terms. Second, we provide conditions under which the target-population LATE is identified using information from the sample population and covariates from the target population. Third, we use two simulation studies to examine how the components of the decomposition change with different forms of misspecification and distribution shift. This allows us to identify when DML's flexible nuisance estimation reduces target-LATE prediction error relative to I2SLS and when substantial error remains.

To be specific, Term I measures error in the estimated sample LATE. Term II captures how error in the estimated treatment-effect slope is amplified when the mean covariates of sample and target compliers differ. Term III captures error from approximating a nonlinear CATE with a linear model when the sample and target complier distributions differ. Term IV captures error from using estimated compliance weights to construct the target complier covariate mean. These four terms are additive but interdependent because errors in the estimated treatment-effect coefficients enter Terms I and II and, through the estimated slope, also affect Term IV.

Identifying the target LATE using information from the sample population and target covariates requires additional assumptions.[3] Conditional on the same observed covariates, compliers must have the same mean treatment effect in the sample and target populations. Also, at each covariate value, the probability of being a complier in the target population must be a constant multiple of that probability in the sample population, which we call proportional compliance. This assumption implies that the overall share of compliers can differ between the two populations, but the way it varies with covariates cannot. Finally, covariate values represented among target compliers must also lie in the support of the

[3]Following Frölich (2007), we express the LATE within a population as a weighted average of covariate-specific complier effects, with weights reflecting the distribution of covariates among compliers.

sample population where there is sufficient instrument variation. Under these conditions, the target LATE can be expressed using the conditional treatment effects and compliance function identified in the sample population together with the target covariate distribution.

We present two simulation exercises to study these channels. First, we construct a controlled Monte Carlo design in which the conditional treatment effect is linear and systematically vary the covariate distribution, the shape of the instrument propensity, the functional form of the additive covariate term in the outcome equation, and the degree of sample-target selection. Because the treatment effect is linear by construction, this exercise lets us isolate how errors in estimating the IV relationship interact with differences between the sample and target complier populations. Second, we use a simulation calibrated to Card (1995), which estimates the return to schooling using proximity to a college as an instrument. In this setting, we separately vary the nonlinearity of the instrument propensity, the relationship between covariates and the outcome, the conditional treatment effect, and the degree of sample-target selection. This allows us to examine, in a more empirically grounded setting, when flexible nuisance estimation improves prediction and when errors remain because the treatment-effect model or the compliance weights do not adequately describe the target complier population.

The relative performance of DML and I2SLS depends on whether functional-form misspecification arises in the instrument propensity, the outcome regression, or the conditional average treatment effect (CATE). When misspecification enters the instrument propensity or the outcome regression, DML can reduce prediction error by flexibly estimating these conditional relationships and using orthogonal moments. In the Card simulation, we vary the nonlinearity of the instrument propensity. With no site selection, a linear CATE, and a target LATE of 0.20, the I2SLS bias is 0.001 when the log odds of the instrument propensity are linear in the covariates. It rises to 0.69 at the propensity calibrated to the Card data and to 1.09 when the nonlinearity is stronger. Over the same settings, the DML bias stays close to zero, between -0.034 and -0.012 (Section 5.5). Likewise, increasing nonlinearity

in the relationship between covariates and the outcome has little effect on DML, whereas I2SLS becomes less accurate. These results show that DML can be less sensitive than I2SLS to misspecification of the nuisance functions.

However, DML's flexible nuisance estimation does not eliminate two sources of prediction error. CATE nonlinearity causes the first. DML's flexibility lies in the nuisance functions. After the nuisance functions are estimated, our DML specification still requires a model for how the treatment effect varies with the covariates, and our baseline model is linear. Therefore, prediction error can remain even under a random sample-target split with no systematic covariate shift.[4] To study this channel, we hold the nuisance estimates and compliance weights fixed and replace the linear treatment-effect model with two cubic basis expansions. The cubic models lower bias in some settings but increase dispersion across replications, illustrating a bias–variance tradeoff in the choice of the treatment-effect model.

Another limitation arises from compliance weighting. Estimating and adjusting the compliance weights can introduce prediction error even when the identification assumptions hold. We estimate the compliance probability as a function of the covariates and evaluate it at the target covariates. However, some target covariate values may be rare or absent in the realized sample even when they lie in the support of the sample population. At those values, the estimated compliance probability is an extrapolation of the model. To avoid this extrapolation, we restrict the evaluation to an interval in which the realized sample provides enough observations and move target covariates outside the interval to its nearest boundary. This restriction is conservative, and the errors it introduces contribute to Term IV. Appendix D.4 examines prediction sensitivity to the choice of interval. DML and I2SLS use the same weights, so errors in these weights can reduce prediction accuracy for both methods. More fundamentally, when proportional compliance or the support condition fails, information from the sample population and target covariates alone need not identify the target

[4]The linear treatment-effect stage of our DML estimator is a weighted linear projection of the CATE. Its projection weights differ from the compliance weights that define the LATE, so a compliance-weighted average of the estimated CATE need not equal the LATE when the CATE is nonlinear.

complier distribution. Our simulations impose the same compliance function in the sample and target populations and therefore do not capture errors that arise when the shape of the compliance function differs across populations.

Our paper relates to the literature on external validity, including Hotz et al. (2005), Galiani and Quistorff (2024), and Allcott (2015), who predict or extrapolate average treatment effects across populations using randomized experiments. In contrast, we study the prediction of the target LATE in an IV setting. Rudolph and van der Laan (2017) study a similar problem, the transport of a complier effect from a sample site to a target site. Their setting requires target-site data on the instrument and treatment. Our setting requires only covariate data from the target population, a less demanding data requirement. We also decompose the prediction error of each estimator into the four terms of Proposition 2 and compare I2SLS with DML.

On IV interpretation, Blandhol et al. (2026) characterize when 2SLS with covariate controls estimates an average of complier effects with nonnegative weights, an estimand they describe as weakly causal. Up to differences in covariate centering, Zhao et al. (2025) analyze the same specification that we use for I2SLS. They derive conditions under which the treatment coefficient recovers the LATE and the interaction coefficients recover the complier-population linear projection of treatment effects on covariates. Mogstad et al. (2018) bound treatment parameters by extrapolating along the latent-resistance margin within a single population. These studies characterize what IV estimators recover within a single population. We build on this literature by studying how an interacted 2SLS estimate can be used to predict a LATE in another population. Moving across populations introduces additional errors associated with differences in complier composition and with the weighting of the target population, which our decomposition makes explicit.

On extrapolation across IV settings, Angrist and Fernandez-Val (2013) reconcile LATEs generated by different instruments within one population under an assumption they call Conditional Effect Ignorability. The common logic is to estimate conditional complier effects

for one group and reweight them using another group's covariate distribution. Their groups are compliers induced by different instruments within the same population. Our groups are compliers for the same instrument in two different populations. The two identifying conditions are therefore distinct, and neither implies the other. Conditional complier effects may differ across instruments within one population even when they are stable across populations for the same instrument, and the reverse may also occur. Some covariate profiles observed in the target population may be rare or absent in the realized sample, even when they occur in the sample population. Predicting treatment effects for those profiles therefore requires functional-form extrapolation. In our baseline comparison, both I2SLS and DML use a linear CATE model, so DML's flexible nuisance estimation does not relax this restriction. In the Card simulation, we separately examine DML specifications that allow the estimated CATE to be nonlinear in the covariates.

In the machine-learning literature on causal inference, Chernozhukov et al. (2018) develop the orthogonalization and cross-fitting framework underlying our DML estimator.[5] Syrgkanis et al. (2019) develop related machine-learning estimators for heterogeneous treatment effects in IV settings. Both focus on estimation within a single population. We study how DML performs when its estimated treatment-effect relationship is used to predict a LATE in another population, and compare its performance with I2SLS.

The remainder of the paper is organized as follows. Section 2 introduces the framework, assumptions, and estimators. Section 3 presents the bias decomposition. Sections 4 and 5 report the Monte Carlo and the Card-calibrated simulation results. Section 6 concludes. Appendix C reports the cooking-fuel application, which Section 1.1 below introduces.

[5]We implement DML using the `OrthoIV` estimator in `econml` (Battocchi et al., 2019). Chernozhukov et al. (2018) state the residualized IV score for a constant treatment effect, whereas `OrthoIV` estimates a heterogeneous treatment-effect function.

1.1 Motivating Example

We illustrate the prediction problem with an application to solid fuel use and cooking time following Kwon and Lee (2025). The treatment is whether a household cooks with solid fuel, and the outcome is the household's average daily cooking time in hours. The instrument indicates whether the share of other households in the same neighborhood using solid fuel is above the country average, a proxy for availability. For each country, we predict its LATE using data from the other countries and compare the prediction with the LATE estimated by 2SLS from that country's own data.

The countries are Ethiopia, Honduras, Kenya, Cambodia, Nepal, and Zambia. The data are from the World Bank's Multi-Tier Framework survey, which provides nationally representative samples with harmonized variables across countries. For each target country, the sample population pools the other five countries, so each country serves once as the target and five times as one of the sample population countries. The application follows the data construction and leave-one-country-out design of Kwon and Lee (2025), and the estimates reported here were re-estimated for this paper.[6]

We compare two predictions. The I2SLS prediction, explained in more detail in Section 2.5, estimates a linear interacted IV model to the five-country sample and evaluates it at the target country's covariate mean. For DML, we estimate the treatment-effect relationship using data from the other five countries and average the estimated effects over the target country's households. For DML, we select the best-performing DML specification among the candidate hyperparameter settings by minimizing the distance between its predicted LATE and the target LATE estimated from the country's own data. The target LATE is obtained from a separate 2SLS regression with linear controls. The DML model itself is estimated using data from the other five countries, so the resulting value remains a prediction for a held-out target country. Both I2SLS and DML average their predicted treatment effects over

[6]Nepal has a first-stage F statistic of 9.4, while the values for the other five countries range from 35.5 to 104.7. Nepal's much smaller value warrants particular caution given the finite-sample problems associated with weak instruments (Staiger and Stock, 1997).

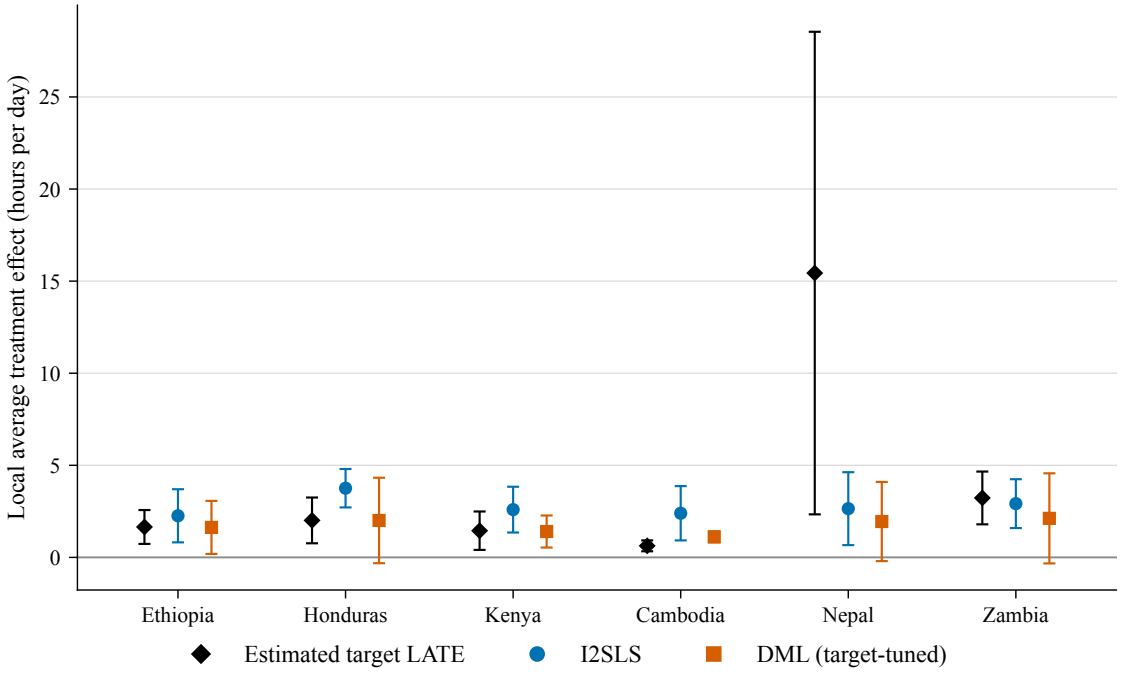


Figure 1: Predicted and Estimated Target LATEs in Six Countries

Notes. The figure compares predicted and estimated effects of solid-fuel use on average daily cooking time, measured in hours. All estimates were recomputed for this paper. Points are estimates, and bars are 95% intervals. The three series report the target LATE estimated from each country's own data, the I2SLS prediction, and the target-tuned DML prediction. The target LATE is estimated separately by 2SLS with linear controls. For each country, the DML hyperparameters were selected to bring the predicted value close to that country's own 2SLS estimate. The DML model is estimated using the other five countries, with hyperparameters selected using the target country's estimated LATE. Its bootstrap interval is based on 100 replications and treats the selected hyperparameters as fixed, so it does not account for the uncertainty introduced by target-based tuning. The I2SLS prediction uses no target outcome, treatment, or instrument data. The target LATE intervals use heteroskedasticity-robust 2SLS standard errors, while the I2SLS intervals use the delta method. In Cambodia, the intervals for the estimated target LATE and the DML prediction are shorter than their markers. Both methods average predicted effects over the target population without applying the compliance weights defined in Section 2.3. Table 4 reports the underlying values.

the full target population using its observed covariates.

The target-tuned exercise shows that DML can improve on the linear I2SLS prediction for some target countries, but the improvement is not uniform. Kwon and Lee (2025) document this uneven cross-country performance but do not explain why prediction works better for some countries than for others. The present paper provides a framework for understanding these differences. It distinguishes errors inherited from estimating the LATE in the original sample from errors that arise when the estimated relationship is used to predict the LATE

for a target population with a different mix of compliers. Nepal illustrates why additional flexibility is not always sufficient. Its estimated target LATE is 15.44 hours per day, while the target-tuned DML prediction is 1.95 and the I2SLS prediction is 2.65. The target-based search in Kwon and Lee (2025) did not find a DML specification closer than 1.95, although it did not exhaust all possible hyperparameter choices. Nepal's estimated target LATE is also the least precise of the six and is based on the weakest first stage. Excluding Nepal from the pooled sample produces little change in the predictions for the other five countries. Appendix C and Table 4 report the full results and discuss the candidate explanations.

The present paper uses this uneven performance as motivation for studying when DML improves on linear I2SLS and why prediction errors remain in some settings. We combine the error decomposition with controlled simulations to distinguish errors that arise when estimating the treatment-effect relationship in the original sample from errors that arise when applying that relationship to a target population with a different mix of compliers.

2 Setup and Estimators

2.1 Framework and Assumptions

Let i index individuals and $D_i \in \{0, 1\}$ indicate whether i belongs to the sample population ($D_i = 1$) or the target population ($D_i = 0$). Each individual has covariates $X_i \in \mathbb{R}^d$, a binary instrument Z_i, and a binary treatment T_i generated by a LATE framework. Let $Y_i(t)$ denote the potential outcome under treatment $t \in \{0, 1\}$, and let $T_i(z)$ denote treatment status when $Z_i = z$. Compliers satisfy $T_i(1) > T_i(0)$, and $C_i = 1\{T_i(1) > T_i(0)\}$ denotes the complier indicator. Always-takers satisfy $T_i(1) = T_i(0) = 1$, never-takers satisfy $T_i(1) = T_i(0) = 0$, and defiers satisfy $T_i(1) < T_i(0)$. We observe N_S independent draws of (Y_i, T_i, Z_i, X_i) from the sample population, which we call the realized sample, and N_T independent draws of X_i from the target population. Write $\mathbb{E}_S[A_i] = \mathbb{E}[A_i \mid D_i = 1]$ and $\mathbb{E}_T[A_i] = \mathbb{E}[A_i \mid D_i = 0]$ for expectations over the two populations.

Define the individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$. Let $\pi_S(x) = \Pr(C_i = 1 \mid X_i = x, D_i = 1)$ and $\pi_T(x) = \Pr(C_i = 1 \mid X_i = x, D_i = 0)$ denote the conditional complier shares. Each population has its own conditional average treatment effect (CATE) among compliers:

$$\tau_S(x) = \mathbb{E}[\tau_i \mid X_i = x,\ T_i(1) > T_i(0),\ D_i = 1], \tag{1}$$

$$\tau_T(x) = \mathbb{E}[\tau_i \mid X_i = x,\ T_i(1) > T_i(0),\ D_i = 0]. \tag{2}$$

Let F_S and F_T denote the covariate distributions among sample and target compliers. The objects of interest are the two LATEs, each defined as a property of its own population:

$$\tau_S = \mathbb{E}[\tau_i \mid D_i = 1,\ T_i(1) > T_i(0)] = \int \tau_S(x)\, dF_S(x) \quad \text{(sample LATE)}, \tag{3}$$

$$\tau_T = \mathbb{E}[\tau_i \mid D_i = 0,\ T_i(1) > T_i(0)] = \int \tau_T(x)\, dF_T(x) \quad \text{(target LATE)}. \tag{4}$$

Throughout, subscripts S and T denote the sample and target populations, respectively. The target LATE can also be written as a compliance-weighted average over the whole target population,

$$\tau_T = \frac{\mathbb{E}_T[\pi_T(X_i)\,\tau_T(X_i)]}{\mathbb{E}_T[\pi_T(X_i)]}. \tag{5}$$

Frölich (2007) shows that, under conditional IV assumptions and a finite outcome mean, the LATE in a single population is identified by comparing units with $Z = 1$ and $Z = 0$ that have the same covariates. At each covariate value, the difference in average outcomes between these units equals the share of compliers times their average treatment effect, and the difference in treatment probabilities equals the share of compliers.[7] Averaging both differences over the covariate distribution and dividing the first by the second gives the

[7] The argument uses the three conditions of Assumption 1 below. Fix a covariate value x within one population and write $\pi(x)$ and $\tau(x)$ for the complier share and the conditional LATE at x. Monotonicity rules out defiers, so the treated units are always-takers and compliers when $Z = 1$ and always-takers alone when $Z = 0$. Because Z is independent of the potential outcomes and compliance types given X, both instrument groups contain the same shares of each type, so $\mathbb{E}[T \mid X = x, Z = 1] - \mathbb{E}[T \mid X = x, Z = 0] = \pi(x)$. Always-takers are treated in both groups and never-takers in neither, and by the exclusion restriction their outcomes do not depend on Z, so their contributions to the outcome difference cancel. Only compliers change treatment with Z, which gives $\mathbb{E}[Y \mid X = x, Z = 1] - \mathbb{E}[Y \mid X = x, Z = 0] = \pi(x)\tau(x)$.

compliance-weighted average in equation (5), in which each covariate value contributes in proportion to its compliers. Angrist and Fernandez-Val (2013) follow the same representation. The goal is to predict τ_T with the estimated $\tau_S(x)$ and $\pi_S(x)$ and with the covariates of the target population. The conditional IV assumptions identify $\tau_S(x)$ and $\pi_S(x)$ in the sample population and do not identify $\tau_T(x)$ or $\pi_T(x)$.

We adapt the conditional IV framework of Frölich (2007) to predict a target-population LATE using estimates from the sample population and target covariates. Assumption 1 states the standard conditional IV restrictions, while Assumptions 2–5 impose the additional conditions sufficient to identify the target LATE when only covariates are observed for the target population.

Assumption 1 (Conditional IV Validity)**.** *Within each population, Z_i is independent of $(Y_i(0), Y_i(1), T_i(0), T_i(1))$ conditional on X_i. The exclusion restriction $Y_i(t,z) = Y_i(t)$ holds for all t, z, and monotonicity $T_i(1) \geq T_i(0)$ holds for all i. Let $h_0(x) = \Pr(Z_i = 1 \mid X_i = x, D_i = 1)$ denote the instrument propensity in the sample population. The subscript 0 indicates that h_0 is the true instrument propensity. For F_T-almost every x, the sample population satisfies $0 < h_0(x) < 1$ and $\pi_S(x) > 0$. The target population has a positive complier share, $\mathbb{E}_T[\pi_T(X_i)] > 0$.*

Assumption 1 requires instrument variation and compliers in the sample population where the target LATE places positive weight. It does not require positive compliance at every x.

The assumption relating treatment effects across the two populations is:

Assumption 2 (Equal Conditional Mean Effect)**.** *For almost every target covariate value with $\pi_T(x) > 0$,*

$$\tau_T(x) = \tau_S(x). \tag{6}$$

Assumption 2 restricts conditional means and not individual effects.[8] Conditional on observed covariates, the mean treatment effect among compliers does not depend on whether

[8]The stronger condition $\tau_i \perp D_i \mid X_i,\ C_i = 1$ implies equation (6). It is sufficient and not necessary.

they belong to the sample or target population. We write $\tau^*(x) \equiv \tau_S(x)$ for the conditional complier effect where Assumption 2 applies. Compliance is proportional across the two populations:

Assumption 3 (Proportional Compliance)**.** *There is a constant $c > 0$ such that*

$$\pi_T(x) = c\,\pi_S(x) \tag{7}$$

for almost every target x, with $0 \leq c\,\pi_S(x) \leq 1$.

Overall compliance rates may differ across the two populations. What Assumption 3 restricts is how compliance varies with X. Equality of the two functions is the case $c = 1$. Without knowing c, the assumption does not identify the target complier share $c\,\mathbb{E}_T[\pi_S(X_i)]$.

The preceding assumptions relate conditional treatment effects and compliance probabilities across the two populations. To use these relationships to identify the target LATE, the conditional treatment effects and compliance probabilities must be identified in the sample population at the relevant target covariate values. We therefore impose the following support and instrument overlap conditions:

Assumption 4 (Support and Instrument Overlap)**.** *In the sample population, $\pi_S(x)$ is identified for almost every x with respect to the target-population distribution of X_i. For almost every such x with $\pi_T(x) > 0$, $\tau_S(x)$ is also identified. Such covariate values lie in the support of the sample-population distribution of X_i and satisfy $0 < h_0(x) < 1$ and $\pi_S(x) > 0$.*

An estimated compliance model may still produce predictions at covariate values outside the support of the sample population, or at values where the instrument takes only one value. At those values, the prediction relies on the model's assumptions rather than an observed comparison between the two instrument groups.

Assumption 5 (Regularity)**.** *The expectations displayed below are finite and $\mathbb{E}_T[\pi_S(X_i)] > 0$.*

Under Assumptions 1 and 4, $\tau_S(x)$ is identified in the sample population at every covariate value represented among target compliers, except possibly on a set of values with probability zero under F_T. Assumption 2 allows us to use these conditional effects for target compliers with the same covariates. Substituting equation (6) into equation (4) therefore expresses the target LATE as an average of conditional complier effects from the sample population:

$$\tau_T = \int \tau^*(x)\, dF_T(x). \tag{8}$$

The estimators considered below estimate $\tau_S(\cdot)$ and use these conditional effects to predict τ_T. Implementing equation (8) also requires F_T, the covariate distribution of target compliers. This distribution is not directly observed because complier status is unobserved. Section 2.2 expresses the same parameter using the observed target covariate distribution.

Assumption 6 imposes a linear specification for the conditional complier effect, allowing extrapolation beyond the covariate range observed in the sample population. We adopt this assumption as a practical functional-form restriction for the I2SLS benchmark. Both I2SLS and OrthoIV with a linear treatment-effect specification rely on this structure.

Both I2SLS and our baseline OrthoIV specification use the following linear model for the conditional complier effect:

Assumption 6 (Linearity in conditional treatment effects)**.** $\tau^*(x) = \alpha_0 + \alpha^\top x$ *for some* $(\alpha_0, \alpha) \in \mathbb{R}^{1+k}$.

We measure each estimator's total prediction error relative to the true target LATE. To decompose this error into the terms of Proposition 2, we compare the estimated coefficients with reference coefficients. When the true conditional complier effect is linear in the covariates, the reference coefficients are its true coefficients. When the effect is nonlinear, it cannot be represented by a single vector of constant slopes, so we use the coefficients of a linear approximation to the true effect. The difference between the estimated coefficients and the reference coefficients captures coefficient estimation error, while the deviation of the

true effect from its linear approximation captures error due to nonlinearity. The choice of approximation affects how the total error is allocated between Term II and Term III, but leaves the total error unchanged.

In the Monte Carlo design, the conditional complier effect is linear, so we use its true coefficients. In the Card simulation, we use the population least-squares projection of the true conditional complier effect onto an intercept and the two covariates, weighted by the complier distribution in a fixed reference population defined before the sample–target split. Because this projection is computed from the known simulation functions, the reference coefficients remain fixed across replications and degrees of site selection. Section 3.1 provides further details.

Because complier status is not directly observed, the sample and target complier covariate means must also be estimated. Section 2.3 explains how we construct the target complier mean used to calculate the predictions.

2.2 Identification of the Target LATE

The next result applies the compliance-weighted representation of Frölich (2007) across two populations. Under Assumption 2, the target's conditional complier effects can be replaced by those of the sample population, and under Assumption 3, its complier probabilities can be replaced by those of the sample population up to a constant factor.

Proposition 1 (Target LATE under proportional compliance)**.** *Under Assumptions 1–5,*

$$\tau_T = \frac{\mathbb{E}_T[\pi_S(X_i)\,\tau_S(X_i)]}{\mathbb{E}_T[\pi_S(X_i)]}. \tag{9}$$

To derive Proposition 1, note that C_i indicates complier status. The target LATE can therefore be written as

$$\tau_T = \mathbb{E}_T[\tau_i \mid C_i = 1] = \frac{\mathbb{E}_T[C_i\tau_i]}{\mathbb{E}_T[C_i]} = \frac{\mathbb{E}_T[\pi_T(X_i)\tau_T(X_i)]}{\mathbb{E}_T[\pi_T(X_i)]}.$$

Under Assumption 2, $\tau_T(X_i) = \tau_S(X_i)$ wherever $\pi_T(X_i) > 0$, which are the only covariate values that receive weight. Substituting the proportional compliance condition $\pi_T(X_i) = c\,\pi_S(X_i)$ gives

$$\tau_T = \frac{\mathbb{E}_T[c\,\pi_S(X_i)\tau_S(X_i)]}{\mathbb{E}_T[c\,\pi_S(X_i)]} = \frac{\mathbb{E}_T[\pi_S(X_i)\tau_S(X_i)]}{\mathbb{E}_T[\pi_S(X_i)]}.$$

The constant $c > 0$ cancels, yielding equation (9).

Equation (9) expresses the target LATE in equation (8) using the covariate distribution of the target population, together with the conditional complier effects and the compliance function of the sample population. Both functions are identified under Assumptions 1 and 4.

Proposition 1 is a population result. To use it, we replace π_S and τ_S with estimates from the sample population and the target expectation with an average over target observations,

$$\widehat{\tau}_T = \frac{\sum_{i:D_i=0} \widehat{\pi}_i\,\widehat{\tau}_S(X_i)}{\sum_{i:D_i=0} \widehat{\pi}_i}. \tag{10}$$

This prediction converges to the target LATE when the averages in its numerator and denominator converge to $\mathbb{E}_T[\pi_S(X_i)\tau_S(X_i)]$ and $\mathbb{E}_T[\pi_S(X_i)] > 0$. Section 2.3 describes how we construct the weights $\widehat{\pi}_i$ and why these conditions can fail.

2.3 Compliance Weights

In equation (10), $\widehat{\tau}_S(x)$ estimates the conditional complier effect $\tau_S(x)$, and $\widehat{\pi}_i$ estimates the compliance probability $\pi_S(X_i)$. We first describe how we construct the compliance weights. The following subsections describe the treatment-effect estimators.

To construct the weights, we estimate treatment probabilities separately for $Z = 1$ and $Z = 0$ in the sample,

$$\widehat{m}_z(x) = \widehat{\mathbb{E}}[T_i \mid Z_i = z, X_i = x, D_i = 1], \qquad z \in \{0, 1\}, \tag{11}$$

and take their difference. Even when monotonicity holds in the population, estimating the

two treatment probabilities separately does not ensure that $\widehat{m}_1(x) \geq \widehat{m}_0(x)$ at every covariate value. Their difference, the estimated complier probability, can therefore be negative. We model each $\widehat{m}_z$ as a scaled logit,

$$\widehat{m}_z(x) = c_z \, \text{logit}^{-1}(a_z + b_z^\top x). \tag{12}$$

where $\text{logit}^{-1}(u) = 1/(1 + e^{-u})$. The parameters are estimated by maximum likelihood separately for $z = 1$ and $z = 0$, with the ceiling c_z restricted to $[0.02, 1]$ and each slope in b_z restricted to $[-50, 50]$.

We specify lower and upper limits for each covariate. Before calculating the difference in treatment probabilities, we move any target covariate value outside these limits to the nearest boundary. The vectors L and U collect these limits. Sections 4 and 5 specify the limits used in each design. We then set any negative probability difference to zero. The adjusted covariates and resulting weights are

$$X_i^* = \min\{U, \max\{L, X_i\}\} \text{ (coordinatewise)}, \qquad \widehat{\pi}_i = \max\{\widehat{m}_1(X_i^*) - \widehat{m}_0(X_i^*), \, 0\}. \tag{13}$$

We use $\widehat{\pi}_i$ as the implemented estimate of $\pi_S(X_i)$ for target observation i.

We use these weights to predict the target LATE using equation (10). Adjusting covariate values to the chosen limits and setting negative treatment-probability differences to zero can change the weights and therefore the predicted LATE. The cutoffs used for these adjustments are choices made when constructing the estimator; they do not follow from the identification assumptions.

The weight construction has three implications for prediction. First, X_i^* is used only to evaluate compliance. The original X_i is used to predict the conditional treatment effect in equation (10). No target observation is deleted, although observations with zero weight do not contribute to the weighted average. Second, using the compliance estimate at the boundary for covariate values beyond it assumes that the boundary value applies there; it

does not identify compliance beyond the chosen limits. Third, negative estimates of complier probabilities can offset positive estimates in the denominator and make the weighted average unstable. We set negative estimates to zero to prevent this cancellation. This adjustment ensures nonnegative weights, but the weights can still differ from the true complier probabilities. Appendix D.3 reports the negativity, effective sample size and boundary diagnostics of these weights in the Monte Carlo design, together with a comparison of alternative covariate limits for calculating compliance weights.

I2SLS and DML use the same weights. Their weighting errors can still differ, because their estimated treatment-effect slopes differ.

Compliance weighting matters when covariates associated with a higher probability of being a complier are also associated with different treatment effects. To measure the importance of compliance weighting, we compare the target LATE with the unweighted average of conditional complier effects over the target population:

$$G = \tau_T - \mathbb{E}_T[\tau_S(X_i)].$$

A positive G means that compliance weighting raises the average effect; a negative G means that it lowers the average. Appendix D reports G for both simulation designs.

2.4 Naive 2SLS

The simplest approach estimates a single LATE from the sample using 2SLS:

$$Y_i = \beta_0 + \alpha_{0,\text{naive}} T_i + \beta^\top X_i + \epsilon_i, \quad \text{instrumented by } Z_i. \tag{14}$$

This ignores treatment effect heterogeneity entirely, applying the same $\hat{\alpha}_{0,\text{naive}}$ to every target population regardless of covariate differences. The naive prediction is $\hat{\tau}_{\text{naive},T} = \hat{\alpha}_{0,\text{naive}}$.

Our main comparison is between I2SLS and DML. We also report naive 2SLS as a simple benchmark that applies the sample estimate directly to the target population, without

adjusting for differences in covariate distributions. This comparison shows whether adjusting for observed treatment-effect heterogeneity improves prediction relative to transferring the sample estimate unchanged.

2.5 Interacted Two-Stage Least Squares (I2SLS)

The interacted specification and the prediction follow Allcott (2015), and Kwon and Lee (2025) estimate a specification of the same form by IV. We estimate from the sample:

$$Y_i = \beta_0 + (\alpha_{0,\text{I2SLS}} + \alpha_{\text{I2SLS}}^\top X_i)T_i + \beta^\top X_i + \epsilon_i. \tag{15}$$

In our implementation both T_i and $T_i X_i$ are endogenous and are instrumented by Z_i and $Z_i X_i$. Because the estimated conditional treatment effect is linear in X, the weighted average in equation (10) can be calculated using the target covariate mean under the estimated compliance weights:

$$\widetilde{X}_T = \frac{\sum_{i:D_i=0} \widehat{\pi}_i X_i}{\sum_{i:D_i=0} \widehat{\pi}_i}. \tag{16}$$

The target LATE prediction is therefore

$$\widehat{\tau}_{\text{I2SLS},T} = \widehat{\alpha}_{0,\text{I2SLS}} + \widehat{\alpha}_{\text{I2SLS}}^\top \widetilde{X}_T. \tag{17}$$

2.6 DML via OrthoIV

Section 4.2 of Chernozhukov et al. (2018) states an orthogonal moment that residualizes the outcome, treatment, and instrument on X, assuming a treatment effect that does not vary with X. Syrgkanis et al. (2019) develop machine-learning estimators of heterogeneous treatment effects with an instrument, using the same residualized moment with a treatment-effect function that varies with X. In our setting, this function is the conditional complier effect $\tau_S(x)$. We estimate the CATE with the OrthoIV estimator in `econml` (Battocchi et al., 2019), which combines nonparametric nuisance estimation with an empirical version of

that moment.

The nuisance functions in the sample population are:

$$q_0(X) = \mathbb{E}_S[Y \mid X], \tag{18}$$

$$p_0(X) = \mathbb{E}_S[T \mid X], \tag{19}$$

$$h_0(X) = \mathbb{E}_S[Z \mid X]. \tag{20}$$

Subtracting these conditional means from the corresponding variables gives the population residuals: $\tilde{Y}_i = Y_i - q_0(X_i)$, $\tilde{T}_i = T_i - p_0(X_i)$, and $\tilde{Z}_i = Z_i - h_0(X_i)$.

Under the conditional IV assumptions, the conditional complier effect satisfies

$$\mathbb{E}_S\left[\left(\tilde{Y}_i - \tau_S(X_i)\tilde{T}_i\right)\tilde{Z}_i \mid X_i\right] = 0. \tag{21}$$

Our baseline OrthoIV specification uses a linear model for the conditional complier effect:

$$\hat{\tau}_{\text{DML}}(x) = \hat{\alpha}_{0,\text{DML}} + \hat{\alpha}_{\text{DML}}^{\top} x.$$

We replace the population residuals in equation (21) with estimated residuals and replace $\tau_S(X_i)$ with the linear effect model. We estimate the coefficients by the method of moments: the residual product has zero sample average, both on its own and when multiplied by each covariate. These conditions do not require the conditional mean to equal zero separately at every covariate value. Section 3.2 explains this distinction.

We predict the target LATE by averaging the estimated conditional complier effects over target observations, using the compliance weights defined in Section 2.3:

$$\hat{\tau}_{\text{DML},T} = \frac{\sum_{i:D_i=0} \widehat{\pi}_i \, \hat{\tau}_{\text{DML}}(X_i)}{\sum_{i:D_i=0} \widehat{\pi}_i}. \tag{22}$$

Both I2SLS and our baseline DML specification use a conditional treatment-effect model

that is linear in X. They differ in how they account for the relationships between covariates and the outcome, treatment, and instrument. The I2SLS specification uses linear covariate controls, whereas DML estimates q_0, p_0, and h_0 using flexible models that allow these relationships to be nonlinear.

Our simulations examine misspecification in the instrument propensity and in the additive covariate term in the outcome equation. Section 3.2 explains how these misspecifications affect the treatment-effect estimates.

3 Bias Decomposition

3.1 Prediction Error Decomposition

We decompose the prediction error of each of the three estimators, its prediction minus the true target LATE, into four additive components. The components compare each estimator's coefficients with reference coefficients α_0 and α, defined as the coefficients of the best linear approximation to the true conditional complier effect. When Assumption 6 holds, the approximation is exact, and α_0 and α are the coefficients in that assumption. When it fails, the effect cannot be represented by a single vector of constant slopes, and the approximation leaves a nonlinear remainder. The difference between the estimated and reference coefficients captures coefficient estimation error, and the remainder captures error due to nonlinearity.[9]

Index the estimators by $k \in \{\text{naive}, \text{I2SLS}, \text{DML}\}$ and write the estimated conditional complier effect as $\widehat{\tau}_k(x) = \widehat{\alpha}_{0,k} + \widehat{\alpha}_k^\top x$. For naive 2SLS, $\widehat{\alpha}_{\text{naive}} = 0$.

Recall that F_S and F_T are the covariate distributions among sample and target compliers, respectively. Their mean covariate vectors are $\bar{X}_{C,S} = \int x \, dF_S(x)$ and $\bar{X}_{C,T} = \int x \, dF_T(x)$. Define $\widehat{\tau}_{k,S} = \widehat{\tau}_k(\bar{X}_{C,S})$ and $\widehat{\tau}_{k,T} = \widehat{\tau}_k(\widetilde{X}_T)$, where $\widetilde{X}_T$ is the target covariate mean calculated using the estimated compliance weights in equation (16).

[9]The choice of approximation changes how the total error is divided between Term II and Term III, but not the total error. Section 5.2 states the reference coefficients used in the Card simulation.

Proposition 2 (Prediction error decomposition)**.** *Under Assumption 2, and with the first moments and residual integrals below finite, the prediction error of estimator k decomposes as:*

$$\begin{aligned}\widehat{\tau}_{k,T} - \tau_T = \quad & \underbrace{(\widehat{\tau}_{k,S} - \tau_S)}_{\text{(I) sample LATE error}} + \underbrace{(\widehat{\alpha}_k - \alpha)^\top (\bar{X}_{C,T} - \bar{X}_{C,S})}_{\text{(II) slope error}\times\text{distribution shift}} \\ & \underbrace{- \int [\tau_S(x) - \alpha_0 - \alpha^\top x]\, d(F_T - F_S)(x)}_{\text{(III) CATE nonlinearity}\times\text{distribution shift}} \\ & + \underbrace{\widehat{\alpha}_k^\top (\widetilde{X}_T - \bar{X}_{C,T})}_{\text{(IV) compliance weighting error}} .\end{aligned} \tag{23}$$

Proof. Assumption 2 and the definitions of F_S and F_T give

$$\tau_T - \tau_S = \alpha^\top (\bar{X}_{C,T} - \bar{X}_{C,S}) + \int [\tau_S(x) - \alpha_0 - \alpha^\top x]\, d(F_T - F_S)(x).$$

Subtract this from $\widehat{\tau}_{k,T} - \widehat{\tau}_{k,S} = \widehat{\alpha}_k^\top (\widetilde{X}_T - \bar{X}_{C,S})$ and add the sample LATE error $\widehat{\tau}_{k,S} - \tau_S$. □

Term I: Sample LATE error. Term I compares the treatment effect predicted by the estimated linear model at the sample complier covariate mean with the true LATE in the sample population, before any adjustment for the target population. Incorrectly modeling the relationship between covariates and the outcome can contribute to this error, as Section 3.2 explains.

Term II: Slope error × distribution shift. Term II is the inner product of the slope estimation error $\widehat{\alpha}_k - \alpha$ and the difference in complier covariate means $\bar{X}_{C,T} - \bar{X}_{C,S}$. By the Cauchy–Schwarz inequality,

$$|\text{Term II}| \leq \|\widehat{\alpha}_k - \alpha\| \cdot \|\bar{X}_{C,T} - \bar{X}_{C,S}\|.$$

The bound has two factors: the magnitude of the slope estimation error and the distance between the sample and target complier covariate means. A larger distance raises the bound

for a given slope error, while more accurate slope estimates lower it. If the slope is estimated without error, Term II is zero regardless of the distance between the two means.

Term III: CATE nonlinearity × distribution shift. Under Assumption 6, $\tau_S(x) = \alpha_0 + \alpha^\top x$, so Term III is zero regardless of differences between the sample and target complier covariate distributions. We relax this assumption in the Card simulation of Section 5. Appendix A describes this term when the conditional complier effect is nonlinear.

Term IV: Compliance weighting error. Term IV is the inner product of the estimated slope and the difference between the estimated weighted target covariate mean $\widetilde{X}_T$, defined in equation (16), and the target complier mean.

Write $\pi_i = \pi_S(X_i)$ and let $\widetilde{X}_{\pi,T} = \sum_{i:D_i=0} \pi_i X_i / \sum_{i:D_i=0} \pi_i$ be the target covariate mean calculated using the true relative compliance weights. When both denominators are positive,

$$\widetilde{X}_T - \bar{X}_{C,T} = \frac{\sum_{i:D_i=0}(\widehat{\pi}_i - \pi_i)(X_i - \widetilde{X}_{\pi,T})}{\sum_{i:D_i=0} \widehat{\pi}_i} + (\widetilde{X}_{\pi,T} - \bar{X}_{C,T}). \tag{24}$$

The first component measures how using estimated weights instead of the true weights changes the target covariate mean. This component is zero if the estimated weights equal the true weights or differ from them only by a common positive scaling factor, which cancels in the weighted mean. The second component arises because the covariate distribution in a finite target sample can differ from that of the target population because of random sampling. Under random sampling, finite weighted first moments, and a positive population mean weight, this component converges to zero as the number of target observations increases. By the Cauchy–Schwarz inequality,

$$|\text{Term IV}| \leq \|\widehat{\alpha}_k\| \cdot \|\widetilde{X}_T - \bar{X}_{C,T}\|.$$

I2SLS and DML use the same processed weights, but Term IV can differ between them

because their estimated treatment-effect slopes differ.

3.2 Sources of Prediction Error in I2SLS and DML

This section explains why I2SLS and DML can produce different treatment-effect estimates, and why prediction error can remain under both methods. We first examine how a nonlinear instrument propensity can contribute to prediction error in I2SLS. We then explain how DML can reduce errors arising from nonlinear relationships between covariates and the instrument or the outcome. Finally, we consider the prediction error that can remain when the conditional complier effect is nonlinear but DML uses a linear treatment-effect model.

Nonlinear instrument propensity in I2SLS. Blandhol et al. (2026) examine when IV estimates with linear covariate controls admit a causal interpretation. Their analysis highlights the role of whether the instrument propensity is captured by those controls. We apply this logic to our interacted 2SLS specification to explain how errors in the treatment-effect coefficients enter the prediction-error decomposition. Because our specification uses both Z_i and its interactions with covariates as instruments, we consider the conditional mean remaining after linear covariate adjustment for each instrument.

Let $\mathbb{L}_S[Z_i \mid X_i = x]$ denote the best linear approximation to the instrument propensity in the sample population, including an intercept. Define the difference between the true instrument propensity and this approximation as

$$\delta(x) = \mathbb{E}_S[Z_i \mid X_i = x] - \mathbb{L}_S[Z_i \mid X_i = x].$$

This distinction between the conditional instrument mean and its linear approximation is central to the analysis of linear IV in Blandhol et al. (2026).

We use the Frisch–Waugh–Lovell theorem for IV regression to express I2SLS after removing its linear covariate controls. Regressing the outcome, each endogenous regressor, and each excluded instrument separately on an intercept and X_i, and then applying 2SLS

to their residuals, gives the same treatment-effect coefficient estimates as the original specification. For I2SLS, this requires residualizing Y_i, T_i, each component of $T_i X_i$, Z_i, and each component of $Z_i X_i$ separately.

For the population analysis below, let $\phi(x) = (1, x^\top)^\top$. The component $\phi_0(x) = 1$ corresponds to the intercept. For each instrument $Z_i \phi_j(X_i)$, define

$$\check{Z}_{ij} = Z_i \phi_j(X_i) - \mathbb{L}_S[Z_i \phi_j(X_i) \mid X_i], \qquad \delta_j(x) = \mathbb{E}_S[\check{Z}_{ij} \mid X_i = x].$$

Here $\mathbb{L}_S[Z_i \phi_j(X_i) \mid X_i]$ is the best linear approximation to the instrument using an intercept and X_i in the sample population. Thus, $\check{Z}_{ij}$ is the population projection residual, and $\delta_j(x)$ is its conditional mean. For $j = 0$, this gives $\delta_0(x) = \delta(x)$.

Because each instrument $Z_i \phi_j(X_i)$ is projected onto a constant and X_i in the sample population, the residual satisfies

$$\mathbb{E}_S[\check{Z}_{ij}] = 0, \qquad \mathbb{E}_S[X_i \check{Z}_{ij}] = 0.$$

Consequently, $\mathbb{E}_S[\delta_j(X_i)] = \mathbb{E}_S[\check{Z}_{ij}] = 0$. Thus, $\delta_j(X_i)$ is either zero almost surely or takes both positive and negative values with positive probability.

Because $\check{Z}_{ij}$ is orthogonal to a constant and X_i, it has zero expected product with any linear function of X_i. Subtracting the linear projections of the outcome and endogenous regressors therefore leaves their expected products with $\check{Z}_{ij}$ unchanged. We can thus use the original outcome and endogenous regressors in the population moment equations below.

We distinguish the true treatment-effect coefficients from the coefficients selected by the population I2SLS equations. Write $\boldsymbol{\alpha} = (\alpha_0, \alpha^\top)^\top$ for the true coefficients under Assumption 6. Let $\boldsymbol{\alpha}^*_{\text{I2SLS}}$ solve

$$\mathbb{E}_S\big[\check{Z}_{ij}\{Y_i - T_i \phi(X_i)^\top \boldsymbol{\alpha}^*_{\text{I2SLS}}\}\big] = 0 \quad \text{for every } j.$$

The vector $\boldsymbol{\alpha}^*_{\text{I2SLS}}$ describes the coefficients selected by I2SLS in the sample population; it need not equal the true coefficient vector $\boldsymbol{\alpha}$.

Under the conditional IV assumptions in Assumption 1, Assumption 6, finite second moments, no perfect collinearity in the regressions used to remove the controls, and a unique population coefficient vector, we have

$$\mathbb{E}_S[\check{Z}_{ij} T_i \phi(X_i)^\top](\boldsymbol{\alpha}^*_{\text{I2SLS}} - \boldsymbol{\alpha}) = \mathbb{E}_S[\delta_j(X_i)\, \mathbb{E}_S[Y_i - T_i \tau_S(X_i) \mid X_i]]\,. \tag{25}$$

The right-hand side is the average product of two functions. The first, $\delta_j(X_i)$, is the conditional instrument mean remaining after removing linear controls. The second is the conditional mean of the outcome after subtracting $T_i \tau_S(X_i)$. If their average product is nonzero for any instrument, the true treatment-effect coefficients do not solve the population I2SLS equations. Nonlinearity in this conditional mean can affect the average product, but neither instrument-propensity nonlinearity nor nonlinearity in this conditional mean alone guarantees that the product has a nonzero average.

Equation (25) concerns the population coefficients. The prediction error in Proposition 2 instead depends on the estimated coefficients $\widehat{\boldsymbol{\alpha}}$, which also contain sampling error. The coefficient error is therefore the sum of two parts,

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*_{\text{I2SLS}}) + (\boldsymbol{\alpha}^*_{\text{I2SLS}} - \boldsymbol{\alpha}).$$

Under the conditions stated above equation (25), $\widehat{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}^*_{\text{I2SLS}}$, so the first part converges to zero in probability as the realized sample grows. The second part is the population discrepancy in equation (25), which does not vanish.

The coefficient error enters Proposition 2 through Terms I and II. When the true conditional complier effect is linear, Term III is zero, Term I equals $(\widehat{\alpha}_0 - \alpha_0) + (\widehat{\alpha} - \alpha)^\top \bar{X}_{C,S}$, and Term II equals $(\widehat{\alpha} - \alpha)^\top (\bar{X}_{C,T} - \bar{X}_{C,S})$. Both terms are zero when the coefficients are estimated without error. Term IV does not depend on the coefficient error. It multiplies

the estimated slope by the error in the weighted target covariate mean, $\widetilde{X}_T - \bar{X}_{C,T}$, which equation (24) divides into weight-estimation error and sampling error. Term IV can therefore remain even when the coefficients are estimated without error.

Nuisance estimation in DML. Recall that $q_0(x) = \mathbb{E}_S[Y_i \mid X_i = x]$ is the conditional outcome mean, $p_0(x) = \mathbb{E}_S[T_i \mid X_i = x]$ is the treatment probability, and $h_0(x) = \mathbb{E}_S[Z_i \mid X_i = x]$ is the instrument probability in the sample population.

With the true instrument propensity, the DML instrument residual $Z_i - h_0(X_i)$ has mean zero at every covariate value. In contrast, subtracting a linear approximation to the instrument propensity can leave a nonzero conditional mean, as represented by $\delta(x)$ above. We now examine how estimation errors in q_0, p_0, and h_0 affect the moment in equation (21).

Let $W_i = (Y_i, T_i, Z_i, X_i)$. Write τ for a candidate conditional-effect function and $\eta = (q, p, h)$ for candidate conditional-mean functions, with true values $\eta_0 = (q_0, p_0, h_0)$. Define the score function

$$\psi(W_i; \tau, \eta) = \big(Y_i - q(X_i) - \tau(X_i)[T_i - p(X_i)]\big)\big(Z_i - h(X_i)\big). \tag{26}$$

At the true conditional complier effect, the moment satisfies Neyman orthogonality (Chernozhukov et al., 2018):

$$\partial_\eta \mathbb{E}_S[\psi(W_i; \tau_S, \eta)]\big|_{\eta=\eta_0} = 0. \tag{27}$$

Small changes in the nuisance functions around their true values therefore have no first-order effect on $\mathbb{E}_S[\psi(W_i; \tau_S, \eta)]$.

For the linear treatment-effect model, OrthoIV multiplies this score by $\phi(X_i)$ to obtain an equation for each coefficient. Let $\Delta_q = \widehat{q} - q_0$, $\Delta_p = \widehat{p} - p_0$, and $\Delta_h = \widehat{h} - h_0$. At the true conditional effect, the expected moment contributed by these nuisance-estimation errors is

$$\mathbb{E}_S[\phi(X_i)\{\Delta_q(X_i) - \tau_S(X_i)\Delta_p(X_i)\}\Delta_h(X_i)]\,. \tag{28}$$

Thus, errors in the outcome and treatment models enter through their products with the instrument-propensity error.

The expectation treats the estimated functions as fixed and averages over an independent observation. Cross-fitting supplies this separation by estimating the nuisance functions on observations other than those used to evaluate the moment. It does not ensure that the estimation errors are small. The usual DML guarantees require these errors and their products to shrink sufficiently quickly, together with the relevant moment, instrument-overlap, and identification conditions. Persistent model misspecification can leave the product in equation (28) nonzero.

Our baseline OrthoIV specification uses a linear CATE model. When the true CATE is nonlinear, accurate nuisance estimates alone do not ensure an accurate target LATE prediction. The linear approximation can make errors that do not cancel when averaged over target compliers. Appendix B explains this weighting issue. Appendix H examines whether nonlinear CATE specifications improve prediction in the Card simulation.

3.3 Limitation: When Assumptions 2 and 3 Fail

When Assumption 2 fails, conditional complier effects can differ between the sample and target populations, so that $\tau_T(x) \neq \tau_S(x)$. Neither I2SLS nor DML corrects this difference. Retaining the remaining assumptions, even the aggregation of the sample conditional effect with the true weights π_S differs from the target LATE by

$$\frac{\mathbb{E}_T[\pi_S(X_i)\tau_S(X_i)]}{\mathbb{E}_T[\pi_S(X_i)]} - \tau_T = \frac{\mathbb{E}_T[\pi_S(X_i)\{\tau_S(X_i) - \tau_T(X_i)\}]}{\mathbb{E}_T[\pi_S(X_i)]}. \tag{29}$$

Both simulations maintain equal conditional mean effects across the sample and target populations. If this assumption fails, compliers with the same observed covariates can have different mean treatment effects across populations. Predicting target effects using the sample population's conditional effects would then introduce an additional error that Proposition 2

does not include. Our simulations do not evaluate this source of prediction error.

When Assumption 3 fails, even perfectly estimated complier probabilities in the sample population may give the wrong weights for the target population. For example, target compliers may be more concentrated at covariate values with larger treatment effects than the weights from the sample population suggest. Using those weights would then understate the target LATE.

When the conditional effect is linear, this error enters Term IV through the difference between the target covariate mean calculated using sample-population complier probabilities and the true target complier mean. Our simulations maintain proportional compliance and therefore do not evaluate this source of prediction error.

4 Monte Carlo Simulation

4.1 Design

The data-generating process makes complier treatment effect heterogeneity a function of observable covariates only, with parameters chosen to vary the instrument-propensity and slope channels across conditions that range from mild to extreme heterogeneity. Because $\tau^*(x)$ is linear in this design, the population CATE nonlinearity $\times$ distribution shift component is zero in all 60 combinations (Appendix E.1), so this grid does not vary the CATE-nonlinearity channel. Covariates are drawn from $X_i \sim \mathcal{N}(0, \sigma_X^2)$ with $\sigma_X \in \{1, 3, 10\}$. Larger values of σ_X produce greater variation in conditional complier effects. We include $\sigma_X = 10$, the largest value considered, as a stress test. A binary instrument is generated as:

$$Z_i \mid X_i \sim \text{Bernoulli}\big(\text{logit}^{-1}(0.5 + 0.3X_i + \phi X_i^3)\big), \quad \phi \in \{0, 0.06\}, \tag{30}$$

where $\phi = 0$ gives a linear index inside the logistic link and $\phi = 0.06$ adds a cubic term that strengthens instrument-propensity misspecification of the kind documented empirically

by Blandhol et al. (2026). The tables label $\phi = 0$ "linear" and $\phi = 0.06$ "cubic." These labels refer to the index inside the link, not to $\mathbb{E}[Z \mid X]$ itself. Because the link is logistic, $\mathbb{E}[Z \mid X]$ is nonlinear in X even at $\phi = 0$, so a linear approximation does not capture the instrument propensity exactly in any of the design combinations. However, our simulation results suggest that this nonlinearity need not lead to substantial prediction bias: I2SLS performs well when $\sigma_X = 1$ and $\phi = 0$.

We assign each unit to one of three compliance types: complier, always-taker, or never-taker. We first draw $\nu_i \sim \text{Uniform}(0, 5)$ and generate complier status as

$$C_i \mid X_i, \nu_i \sim \text{Bernoulli}\big(0.4 \cdot \text{logit}^{-1}(0.4X_i + \nu_i)\big)\,, \tag{31}$$

where $C_i = 1$ denotes a complier. Let A_i indicate always-taker status. For units with $C_i = 0$, draw

$$A_i \mid X_i, C_i = 0 \sim \text{Bernoulli}\left(\frac{p_{AT}(X_i)}{1 - p_C(X_i)}\right),$$

where $p_C(x) = \Pr(C_i = 1 \mid X_i = x)$ and $p_{AT}(x) = 0.6 \cdot \text{logit}^{-1}(-2x/\sigma_X)$. For units with $C_i = 1$, set $A_i = 0$. Thus, $A_i = 1$ denotes an always-taker, and units with $C_i = 0$ and $A_i = 0$ are never-takers.

Treatment depends on compliance type and instrument assignment. Always-takers take treatment regardless of assignment, compliers take treatment only when assigned, and never-takers do not take treatment. The potential treatment states are therefore $T_i(0) = A_i$ and $T_i(1) = A_i + (1 - A_i)C_i$, giving

$$T_i = A_i + (1 - A_i)C_i Z_i. \tag{32}$$

These states satisfy $T_i(1) \geq T_i(0)$ for every unit, so monotonicity in Assumption 1 holds by construction and there are no defiers.

The always-taker share is 0.30 at every covariate scale. At $\sigma_X = 1, 3, 10$, the complier

shares are approximately 0.34, 0.33, and 0.28, and the never-taker shares are approximately 0.36, 0.37, and 0.42, respectively. The latent variable ν_i also enters the outcome equation below, so an unobserved factor affects both complier status and the outcome level.

The individual treatment effect is linear in X up to a level shift by compliance type:

$$\tau_i = 10X_i + 2 + 5 \cdot \mathbf{1}\{i \text{ is a complier}\} + 8 \cdot \mathbf{1}\{i \text{ is an always-taker}\}, \tag{33}$$

so the complier CATE is $\tau^*(x) = 10x + 7$, linear with slope $\alpha = 10$. The two offsets shift the level of the effect by compliance type and leave the complier slope at 10. The outcome equation is:

$$Y_i = \tau_i T_i + 2\nu_i + g(X_i) + 0.1\varepsilon_i, \quad \varepsilon_i \sim \text{Uniform}(0, 1), \tag{34}$$

where $g(X_i) = 5X_i$ (linear) or $g(X_i) = 5 \cdot \mathbf{1}\{X_i > 0\}$ (step). Under the step specification, $g(X_i)$ is nonlinear in X_i. Using linear controls to approximate this step function can introduce errors in the estimated treatment-effect coefficients, which enter Terms I and II of Proposition 2.

Selection into the target population depends on observables:

$$p_i^{target} = \text{logit}^{-1}(\vartheta X_i), \quad \vartheta \in \{0, 0.25, 0.50, 0.75, 1.0\}, \tag{35}$$

where ϑ controls selection into the target population. At $\vartheta = 0$, assignment is random, so the sample and target populations have the same covariate distribution. At $\vartheta = 1$, units with larger X_i are more likely to enter the target population. In each simulation draw, we evaluate prediction error relative to the true LATE among that draw's target compliers.

We combine three covariate scales, two instrument specifications, and two outcome specifications to obtain 12 designs. We evaluate each design at five selection strengths, giving 60 combinations in total. Each combination is replicated 500 times with $N = 10{,}000$ per draw.

For DML, we estimate the nuisance functions using XGBoost with two-fold cross-fitting.

We divide the sample into two groups, train the nuisance models on one group, and predict the outcome, treatment, and instrument in the other, then reverse the roles of the groups. The resulting predictions are subtracted from the observed values to obtain the residuals used in estimation. We use 300 trees, a maximum tree depth of 3, and L_1 and L_2 penalties of 0 and 1, respectively. These settings remain fixed across all parameter combinations.

For the Monte Carlo simulations, we calculate compliance weights after moving target covariate values outside a chosen interval to the nearest boundary. This interval is the intersection of (i) the range between the 0.5th and 99.5th percentiles of X_i in the sample and (ii) the range of sample covariate values with estimated instrument probabilities between 0.01 and 0.99. At each covariate value, the two populations have the same conditional complier effect and compliance probability. Thus, Assumptions 2 and 3 hold by construction, with $c = 1$.

4.2 Results

Table 1 reports bias and RMSE for sixteen representative combinations, and Figure 2 compares the RMSE of the two adjusted estimators across all 60 combinations.

DML has lower RMSE than I2SLS in most combinations under both specifications of $g(X)$. Its advantage is more common with linear $g(X)$ than with step-function $g(X)$. I2SLS remains competitive when $\sigma_X = 1$: most combinations in which it outperforms DML occur at this scale.

The comparison with naive 2SLS also shows that adjustment does not always improve prediction. Although naive 2SLS has the largest RMSE in most combinations, I2SLS performs worse in some settings with larger values of σ_X. The difference can be substantial. At $\sigma_X = 10$, with the linear instrument index, step-function $g(X)$, and $\vartheta = 1$, the RMSE is 742.805 for I2SLS and 4.965 for DML. Thus, the results favor DML in a wider range of settings, while showing that neither adjusted estimator uniformly outperforms the other.

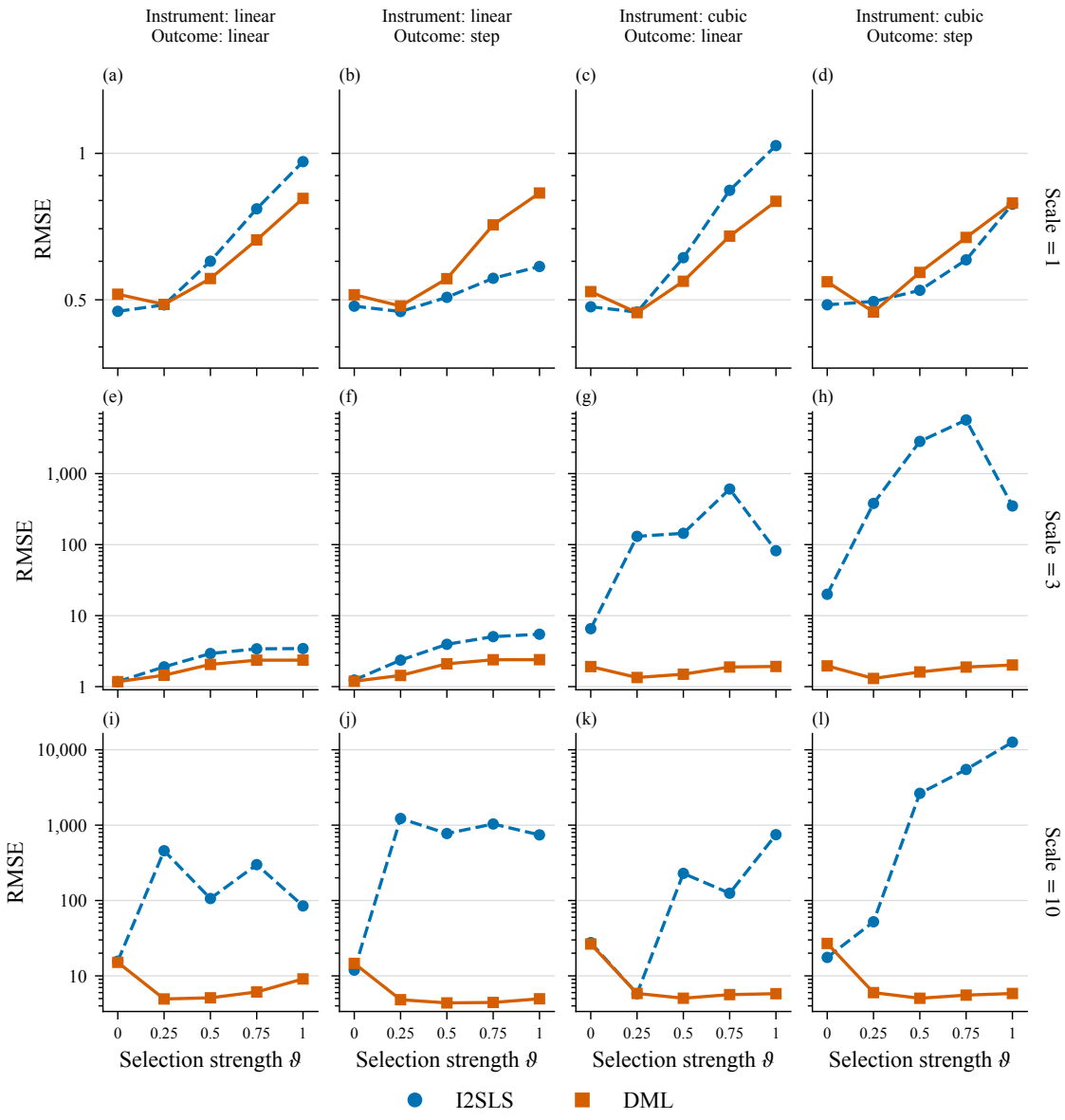


Figure 2: Root Mean Squared Prediction Error across the Full Monte Carlo Grid

Notes. The figure reports RMSE for I2SLS and DML in all 60 design cells. Table 1 and Appendix E.2 report naive 2SLS. Rows are the three covariate scales and columns are the four instrument-by-outcome combinations. The horizontal axis shows the five values of ϑ and the vertical axis is logarithmic.

Table 1: Prediction Bias and RMSE in the Monte Carlo Design

		Naive 2SLS		I2SLS		DML	
Outcome	ϑ	Bias (MCSE)	RMSE	Bias (MCSE)	RMSE	Bias (MCSE)	RMSE
Panel A: scale $= 1$							
Instrument: linear							
linear	0	−0.645 (0.023)	0.828	0.100 (0.021)	0.473	0.164 (0.022)	0.513
	1	−8.590 (0.022)	8.604	−0.788 (0.025)	0.962	−0.601 (0.024)	0.808
step	0	−0.628 (0.025)	0.839	0.109 (0.021)	0.485	0.151 (0.022)	0.512
	1	−8.602 (0.022)	8.615	−0.040 (0.026)	0.585	−0.618 (0.025)	0.829
Instrument: cubic							
linear	0	−0.767 (0.023)	0.928	0.156 (0.021)	0.484	0.186 (0.022)	0.520
	1	−8.264 (0.021)	8.277	−0.868 (0.025)	1.037	−0.590 (0.024)	0.797
step	0	−0.984 (0.024)	1.123	−0.016 (0.022)	0.488	0.215 (0.022)	0.544
	1	−8.567 (0.022)	8.581	0.514 (0.027)	0.786	−0.574 (0.024)	0.791
Panel B: scale $= 10$							
Instrument: linear							
linear	0	−36.379 (0.262)	36.846	−14.792 (0.230)	15.658	−14.142 (0.240)	15.121
	1	−117.343 (0.181)	117.412	−5.312 (3.786)	84.736	1.089 (0.405)	9.104
step	0	−32.482 (0.263)	33.010	−10.766 (0.230)	11.925	−13.726 (0.228)	14.641
	1	−115.837 (0.179)	115.906	65.295 (33.124)	742.805	−0.269 (0.222)	4.965
Instrument: cubic							
linear	0	−41.850 (0.374)	42.676	−27.125 (0.225)	27.588	−25.918 (0.239)	26.461
	1	−111.026 (0.223)	111.138	28.747 (33.443)	747.619	1.489 (0.251)	5.801
step	0	−32.465 (0.370)	33.501	−16.769 (0.236)	17.576	−26.422 (0.233)	26.931
	1	−106.408 (0.230)	106.533	95.143 (565.143)	12624.697	−0.198 (0.261)	5.837

Notes. The table reports sixteen of the 60 cells, using covariate scales 1 and 10, $\vartheta \in \{0, 1\}$ and all four instrument-by-outcome combinations. Appendix E.2 reports the full grid. Each cell has 500 replications of $N = 10{,}000$. Bias is mean prediction error relative to the true LATE of the realized target compliers, parentheses report its Monte Carlo standard error, and RMSE is root mean squared prediction error. Estimators, common weights and the covariate limits used to calculate compliance weights are defined in Section 4.1. Instrument labels refer to the index inside the logistic link.

4.3 Term Decomposition

Table 2 decomposes the prediction error into the four components of Proposition 2 for the same sixteen combinations.

Term III is zero because the conditional complier effect is linear by construction. Prediction error therefore comes from Terms I, II, and IV.

The decomposition shows how each component contributes to bias in predicting the LATE in the target population and which components are smaller for DML than for I2SLS.

Table 2: Prediction Error Components in the Monte Carlo Design

Outcome	ϑ	Bias	Term I	Term II	Term III	Term IV
Panel A: I2SLS						
Scale = 1, *instrument: linear*						
linear	0	0.100 (0.021)	−0.021 (0.012)	0.000 (0.001)	0.000 (0.000)	0.120 (0.018)
	1	−0.788 (0.025)	−0.008 (0.012)	−0.113 (0.013)	0.000 (0.000)	−0.667 (0.016)
step	0	0.109 (0.021)	0.029 (0.014)	0.000 (0.001)	0.000 (0.000)	0.079 (0.017)
	1	−0.040 (0.026)	0.036 (0.014)	0.675 (0.015)	0.000 (0.000)	−0.751 (0.019)
Scale = 1, *instrument: cubic*						
linear	0	0.156 (0.021)	0.026 (0.013)	0.000 (0.001)	0.000 (0.000)	0.130 (0.016)
	1	−0.868 (0.025)	0.039 (0.012)	−0.290 (0.017)	0.000 (0.000)	−0.617 (0.016)
step	0	−0.016 (0.022)	−0.200 (0.013)	−0.001 (0.001)	0.000 (0.000)	0.185 (0.018)
	1	0.514 (0.027)	−0.276 (0.013)	1.544 (0.019)	0.000 (0.000)	−0.754 (0.020)
Scale = 10, *instrument: linear*						
linear	0	−14.792 (0.230)	−0.455 (0.046)	0.003 (0.002)	0.000 (0.000)	−14.340 (0.239)
	1	−5.312 (3.786)	−0.074 (0.085)	−4.410 (3.833)	0.000 (0.000)	−0.828 (0.078)
step	0	−10.766 (0.230)	2.734 (0.053)	0.005 (0.004)	0.000 (0.000)	−13.505 (0.220)
	1	65.295 (33.124)	−2.305 (1.461)	69.287 (34.884)	0.000 (0.000)	−1.688 (0.517)
Scale = 10, *instrument: cubic*						
linear	0	−27.125 (0.225)	−1.134 (0.032)	0.003 (0.001)	0.000 (0.000)	−25.994 (0.230)
	1	28.747 (33.443)	−5.680 (5.457)	35.926 (39.292)	0.000 (0.000)	−1.499 (0.419)
step	0	−16.769 (0.236)	9.414 (0.052)	−0.009 (0.002)	0.000 (0.000)	−26.174 (0.221)
	1	95.143 (565.143)	−11.583 (44.157)	105.357 (615.440)	0.000 (0.000)	1.369 (8.705)
Panel B: DML						
Scale = 1, *instrument: linear*						
linear	0	0.164 (0.022)	0.043 (0.014)	0.000 (0.000)	0.000 (0.000)	0.120 (0.018)
	1	−0.601 (0.024)	0.031 (0.013)	0.050 (0.012)	0.000 (0.000)	−0.682 (0.016)
step	0	0.151 (0.022)	0.070 (0.014)	0.001 (0.001)	0.000 (0.000)	0.079 (0.017)
	1	−0.618 (0.025)	0.031 (0.013)	0.048 (0.012)	0.000 (0.000)	−0.697 (0.017)
Scale = 1, *instrument: cubic*						
linear	0	0.186 (0.022)	0.056 (0.015)	0.000 (0.001)	0.000 (0.000)	0.130 (0.016)
	1	−0.590 (0.024)	0.026 (0.013)	0.026 (0.013)	0.000 (0.000)	−0.642 (0.016)
step	0	0.215 (0.022)	0.036 (0.014)	0.000 (0.001)	0.000 (0.000)	0.179 (0.017)
	1	−0.574 (0.024)	0.005 (0.013)	0.058 (0.013)	0.000 (0.000)	−0.637 (0.016)
Scale = 10, *instrument: linear*						
linear	0	−14.142 (0.240)	0.180 (0.033)	−0.001 (0.002)	0.000 (0.000)	−14.322 (0.234)
	1	1.089 (0.405)	−0.350 (0.071)	2.307 (0.452)	0.000 (0.000)	−0.868 (0.083)
step	0	−13.726 (0.228)	0.117 (0.030)	−0.001 (0.002)	0.000 (0.000)	−13.842 (0.225)
	1	−0.269 (0.222)	−0.142 (0.041)	0.811 (0.236)	0.000 (0.000)	−0.938 (0.076)
Scale = 10, *instrument: cubic*						
linear	0	−25.918 (0.239)	0.036 (0.097)	0.006 (0.006)	0.000 (0.000)	−25.959 (0.232)
	1	1.489 (0.251)	−0.662 (0.083)	3.165 (0.309)	0.000 (0.000)	−1.014 (0.077)
step	0	−26.422 (0.233)	−0.119 (0.093)	−0.001 (0.006)	0.000 (0.000)	−26.301 (0.224)
	1	−0.198 (0.261)	−0.279 (0.085)	1.275 (0.321)	0.000 (0.000)	−1.194 (0.084)

Notes. The table reports the same sixteen representative cells as Table 1. Entries are replication means. Parentheses report Monte Carlo standard errors, computed separately for each component as its replication standard deviation (with divisor $R-1$) divided by $\sqrt{R}$. Terms I–IV sum to the prediction error in each replication. Bias is the mean of the stored prediction errors, with its MCSE computed directly. Complier means and LATEs use realized sample and target compliers in each draw. These diagnostics use simulation truth and are not directly observable in an application. The methods share the compliance weights, but their weighting-error components can differ because their estimated slopes differ. Rounding can prevent the displayed components from summing exactly. Terms I–IV are the components of Proposition 2: sample LATE error, slope error × distribution shift, nonlinearity × distribution shift, and compliance weighting error, respectively. Each cell uses $R = 500$ replications. The reference slope α is the true linear CATE slope, so the nonlinearity component is zero up to numerical precision. The adjusted estimators use the common compliance-weighted endpoint defined in Section 4.1.

When $\vartheta = 0$, assignment to the sample or target population is independent of X. The two populations therefore have the same complier covariate distribution, and Term II is close to zero, with small differences arising from finite-sample variation. When $\vartheta = 1$, assignment depends on X, creating differences in complier covariate means between the two populations. These differences allow slope estimation errors to contribute to prediction error. For I2SLS, this contribution is particularly large in some settings with $\sigma_X = 10$, although the large Monte Carlo standard errors indicate that its magnitude is imprecisely estimated.

DML's error in estimating the LATE in the sample population remains small in the reported combinations. Its Term II contributions at $\sigma_X = 10$ and $\vartheta = 1$ are also much smaller than the largest contributions observed for I2SLS. These results illustrate how DML can reduce prediction bias in settings with a nonlinear instrument propensity and nonlinear $g(X)$.

However, both I2SLS and DML remain exposed to compliance weighting error. At $\sigma_X = 10$, with the cubic instrument index and $\vartheta = 0$, Term IV is approximately -26 for both estimators under both outcome specifications. It is the largest component in absolute value for I2SLS in these combinations. Both estimators use the same compliance weights; their Term IV contributions differ only through their estimated treatment-effect slopes. Thus, DML's smaller errors in estimating treatment effects do not eliminate errors from compliance weighting.

5 Card (1995) Calibrated Simulation

The preceding Monte Carlo results may depend on the particular data-generating process we specify. To examine whether the findings extend to a setting informed by observed data, we conduct a simulation calibrated to the Card data. We choose model parameters to match selected moments of these data, so that the simulated data reproduce these observed features.

We use a calibrated simulation because the true LATE in the target population is gener-

ally unobserved in an empirical case study. The simulation provides a known target LATE, allowing us to evaluate prediction errors directly. It also allows us to vary separately the nonlinearity of the conditional complier effect, the instrument propensity, and the relationship between covariates and the outcome, and to examine how these changes affect prediction as the covariate distributions of the sample and target populations diverge.

We vary ϑ to examine how prediction changes as covariate distributions differ between the sample and target populations. The parameters κ and γ_y control the degree of nonlinearity in the instrument propensity and in the relationship between covariates and the outcome, respectively. Higher values introduce stronger nonlinearity. In addition, we use λ to examine prediction when the conditional complier effect is nonlinear.

5.1 DGP Calibration

Following Blandhol et al. (2026), we calibrate the simulation model to match selected moments of the Card (1995) data. We perform this calibration once and keep the resulting reference distribution, probability functions, and outcome coefficients fixed across simulation replications. The calibration uses the Card sample of the Blandhol et al. (2026) replication package, which contains 3,010 observations. The outcome is the log wage in 1976, the treatment T_i is an indicator for at least 13 years of schooling, and the instrument Z_i is an indicator for a four-year college in the county.

The simulation uses two covariates, potential experience and Census division. Potential experience is age in 1976 minus years of schooling minus six. We cap experience at 20 years and divide it by 20 to obtain $X_1 = \min\{\text{experience}, 20\}/20 \in [0, 1]$. We encode the nine Census divisions as $X_2 \in \{0, 1/8, \ldots, 1\}$, using the division-to-value mapping in Blandhol et al. (2026). Appendix G lists this mapping.

The Card data contain 177 distinct (X_1, X_2) combinations. We assign each combination a probability equal to its share of the 3,010 Card observations and call the resulting distribution the reference covariate distribution. These probabilities are fixed during calibration and

across simulation replications.

We construct the instrument probability using the cubic polynomial specification and coefficients supplied in the replication package of Blandhol et al. (2026):

$$\Pr(Z_i = 1 \mid X_i) = \sum_{j=0}^{3} \sum_{k=0}^{3} c_{jk} \, X_{1i}^{j} \, X_{2i}^{k}. \tag{36}$$

This specification allows the relationship between experience and the instrument probability to differ across Census divisions. Appendix G reports the coefficient matrix and its indexing. We set values below 0.01 to 0.01 and values above 0.99 to 0.99, leaving values within this interval unchanged. This adjustment does not change any probabilities at the covariate values used in the calibration, where they range from 0.1000 to 0.9763.

We next construct the compliance probabilities. We first examine differences in treatment rates between the two instrument groups within Census divisions, experience categories, and their joint cells. These comparisons inform a specification in which compliance varies with both experience and region. Rather than using the raw cell differences directly, we estimate treatment-probability models.

For Card observations with at least six years of experience, we estimate separate linear probability models for $Z = 0$ and $Z = 1$, using cubic terms and their interactions in (x_1, x_2). We restrict the predicted probabilities to $[0, 1]$ and denote them by $p_{T|Z=0}(x)$ and $p_{T|Z=1}(x)$. For observations with at most five years of experience ($x_1 \leq 0.25$), all Card observations are treated in both instrument groups. We therefore set both treatment probabilities to one in these cells. The resulting DGP assigns them to the always-taker type. These observations account for 20.299% of the reference population.

The positive difference between these probabilities,

$$P_c(x) = \max\{p_{T|Z=1}(x) - p_{T|Z=0}(x), 0\},$$

provides the initial complier share at each covariate value. The initial treatment-probability

estimates imply an overall complier share of approximately 0.077. With a small complier share, sampling variability can make performance differences between linear and nonlinear settings difficult to distinguish. To make these comparisons more informative at the chosen sample size, we increase the complier probabilities by a common factor:

$$p^{\zeta}_{T|Z=1}(x) = \min\{p_{T|Z=0}(x) + \zeta P_c(x), 1\}. \tag{37}$$

We use $\zeta = 3.0334$, giving an overall complier share of approximately 0.234 in the reference population. The upper bound does not bind in this calibration, so the adjustment preserves the relative pattern of complier probabilities across covariate values. Appendix G describes the rule used to select the scale.

Conditional on $X = x$, the always-taker probability is $p_{T|Z=0}(x)$, the complier probability is $p^{\zeta}_{T|Z=1}(x) - p_{T|Z=0}(x)$, and the never-taker probability is $1 - p^{\zeta}_{T|Z=1}(x)$. These probabilities exclude defiers and therefore satisfy monotonicity. Because the complier probability varies with the covariates, compliers generally have a different covariate distribution from the whole population.

We then calibrate the conditional potential-outcome means. Let K_i denote compliance type. For each type g and treatment status t, we specify

$$\mathbb{E}[Y_i(t) \mid K_i = g, X_i = x] = \varphi(x)^{\top}\theta_{g,t}, \tag{38}$$

where $\varphi(x)$ contains the 16 terms $x_1^j x_2^k$ with $j, k \in \{0, 1, 2, 3\}$. The six combinations of compliance type and treatment status therefore require 96 coefficients.

We select these coefficients using five summary statistics from the Card data, following Blandhol et al. (2026): the mean log wage, $\mathbb{E}[Y] = 6.26$; the OLS treatment coefficient, $\beta_{\text{OLS}} = 0.24$; the 2SLS treatment coefficient, $\beta_{\text{IV}} = 0.66$; the partially linear IV coefficient, $\beta_{\text{rich}} = 0.43$; and the LATE estimate, $\beta_{\text{LATE}} = 0.20$.

The OLS and 2SLS estimates use Card's controls, including potential experience, its

square, and region indicators. Following Blandhol et al.'s notation, β_{rich} is the IV coefficient estimated using DML to control flexibly for covariates. The estimate β_{LATE} also uses DML, but targets the average treatment effect among compliers. Unlike the LATE, β_{rich} additionally weights conditional complier effects by the conditional variance of the instrument, so the two quantities can differ. Each reported DML estimate is the median over 100 random sample splits. The calibration targets are rounded to two decimal places.

For any candidate set of the 96 coefficients, the conditional outcome means change, while the reference covariate distribution, instrument propensity, and compliance probabilities remain fixed. We combine the candidate means across compliance types, instrument states, and the resulting treatment states, using these fixed probabilities and the Card frequency of each of the 177 covariate combinations. This calculation gives the population mean outcome. We obtain the population OLS and 2SLS coefficients from the corresponding population moment equations, using an intercept, X_{1i}, and region indicators as controls. The simulated LATE averages the conditional complier effects, weighting each cell by its Card frequency multiplied by its complier probability and normalizing these weights to sum to one. The simulated β_{rich} additionally weights each cell by $\Pr(Z = 1 \mid X = x)\{1 - \Pr(Z = 1 \mid X = x)\}$, the conditional variance of the binary instrument. These calculations use the fixed 177-cell population, not a randomly generated simulation sample.

We choose the 96 coefficients by constrained optimization, adapting the calibration approach of Blandhol et al. (2026) to our simulation design. The five population quantities enter as equality constraints. The objective is the sum of two squared differences. Under monotonicity, observations with $Z = 0$ and $T = 1$ are always-takers, while observations with $Z = 1$ and $T = 0$ are never-takers. The first difference compares the mean log wage among Card observations with $Z = 0$ and $T = 1$ with the model's treated always-taker mean. We calculate this model mean at each of the 177 covariate combinations and average it using the combination's frequency in the Card data. The second difference compares the mean log wage among Card observations with $Z = 1$ and $T = 0$ with the model's untreated

never-taker mean, also calculated across the 177 covariate combinations using their Card frequencies. These two model averages use only the reference covariate frequencies, without additional weighting by compliance-type probabilities or the instrument propensity.

Of the 16 terms in $\varphi(x)$, the 13 terms other than 1, x_1, and x_2 are the nonlinear basis terms. For each treatment status and each of these 13 terms, an additional equality constraint sets the weighted sum of the three compliance-type coefficients to zero. The weights are the unconditional compliance-type shares in the reference population. Other equality constraints equate the two never-taker potential-outcome coefficient vectors, thereby imposing a zero conditional treatment effect for never-takers. Inequality constraints bound the conditional outcome means between 4.6 and 7.79 and bound each type-specific treatment effect in absolute value by 2. They also require a nonnegative treatment effect for always-takers. The resulting conditional complier effect can be nonlinear in the covariates. Appendix G gives the complete objective and constraints and reports the numerical accuracy of the accepted solution.

The calibrated specification corresponds to $\lambda = 1$, $\kappa = 1$, and $\gamma_y = 0$. It retains the calibrated conditional complier effect and instrument propensity. Although $\gamma_y = 0$, the relationship between the covariates and the outcome can already be nonlinear because equation (38) includes cubic terms and their interactions. Across the parameter grids used below, higher values of λ, κ, and γ_y introduce more nonlinearity in the corresponding component. We apply these parameter changes to the fixed calibrated functions and coefficients. We do not recalibrate the Card moments after changing a parameter.

5.2 Simulation Design

Before running the Monte Carlo replications, we compute the two linear approximations described below using the Card data and the reference population and keep them fixed across all replications.

For each Monte Carlo replication, we generate a population of $N = 18{,}000$ observations

from the fixed DGP. We first draw the (X_{1i}, X_{2i}) pairs jointly with replacement from the 3,010 Card observations. The reference covariate distribution therefore determines the sampling probabilities, while the realized cell shares can differ across replications.

The first approximation is to the calibrated instrument propensity on the log-odds scale. Let $s(X_i) = \text{logit}\big(\mathbb{E}[Z_i \mid X_i]\big)$ denote the calibrated log odds, where $\text{logit}(p) = \log\{p/(1-p)\}$, and let $s_{\text{lin}}(X_i) = \mathbb{L}\big[s(X_i) \mid 1, X_{1i}, X_{2i}\big]$ denote its best linear approximation. We obtain this approximation by regressing $s(X_i)$ on an intercept, X_{1i}, and X_{2i} by OLS using the covariate values observed in the Card data. These regression coefficients are fixed as described above.

The parameter κ scales the difference between the calibrated log odds and this linear approximation:

$$\mathbb{E}_\kappa[Z_i \mid X_i] = \text{logit}^{-1}\Big(s_{\text{lin}}(X_i) + \kappa\big(s(X_i) - s_{\text{lin}}(X_i)\big)\Big), \qquad \kappa \in \{0, 1, 2\}, \tag{39}$$

where $\text{logit}^{-1}(u) = 1/(1 + e^{-u})$. At $\kappa = 0$, we use only the linear approximation on the log-odds scale, although the logistic link can still make the resulting probability nonlinear in X. At $\kappa = 1$, we recover the calibrated instrument probability, and at $\kappa = 2$, we double the departure from the linear approximation on the log-odds scale. In this calibration, the departure from a linear probability model increases with κ.[10] We evaluate equation (39) at each observation's covariates and draw Z_i from a Bernoulli distribution with that probability.

We next draw a single $U_i \sim \text{Uniform}(0, 1)$, independently of (X_i, Z_i). An observation is an always-taker when $U_i \le p_{T|Z=0}(X_i)$, a complier when $p_{T|Z=0}(X_i) < U_i \le p^{\zeta}_{T|Z=1}(X_i)$, and a never-taker otherwise. Using the same draw, treatment is $T_i = \mathbf{1}\{U_i \le p^{\zeta}_{T|Z=1}(X_i)\}$ when $Z_i = 1$ and $T_i = \mathbf{1}\{U_i \le p_{T|Z=0}(X_i)\}$ when $Z_i = 0$.

Write $K_i = \text{C}$ for the complier type, and let

$$\tau^*(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid K_i = \text{C}, X_i = x] = \varphi(x)^\top(\theta_{\text{C},1} - \theta_{\text{C},0})$$

[10] Under the reference covariate distribution, the root mean square error of the linear approximation of $\mathbb{E}_\kappa[Z_i \mid X_i]$ on $(1, X_1, X_2)$ is 0.004343, 0.141367 and 0.242180 at $\kappa = 0$, 1 and 2, respectively.

denote the calibrated conditional complier effect implied by equation (38). We obtain a linear approximation to $\tau^*(x)$ by regressing it on an intercept, X_1, and X_2 using weighted least squares. Each covariate combination is weighted by its frequency in the Card data multiplied by its calibrated complier probability. Denote this projection by $\mathbb{L}[\tau^* \mid X = x]$. This compliance-weighted projection is also fixed as described above.

The parameter λ scales the difference between the calibrated conditional complier effect and this projection:

$$\tau_\lambda(x) = \mathbb{L}[\tau^* \mid X = x] + \lambda \cdot \big(\tau^*(x) - \mathbb{L}[\tau^* \mid X = x]\big), \qquad \lambda \in \{0, 1\}. \tag{40}$$

At $\lambda = 0$, the conditional complier effect equals its compliance-weighted linear approximation. At $\lambda = 1$, it equals the full nonlinear effect function specified in the calibration. The construction preserves the average effect among compliers in the reference population at 0.200.

For each observation, we first initialize $Y_i(t)$ at the conditional mean in equation (38) for $t \in \{0, 1\}$, using that observation's covariates and compliance type. To impose $\tau_\lambda(X_i)$, we leave a complier's initialized value of $Y_i(0)$ unchanged and replace its initialized value of $Y_i(1)$ with $Y_i(0) + \tau_\lambda(X_i)$. The initialized potential outcomes of always-takers and never-takers remain unchanged. We then apply the common covariate-dependent shift to these intermediate values:

$$Y_i(t) \longmapsto Y_i(t) + \gamma_y \mathbf{1}\{X_{1i} > 0.5\}, \qquad t \in \{0, 1\}, \qquad \gamma_y \in \{0, 2, 10\}. \tag{41}$$

Because the shift is the same for both potential outcomes, it changes the relationship between covariates and the outcome without changing the individual treatment effect. The original outcome calibration targets and bounds are not imposed again after this shift. Finally, we draw one error from a normal distribution with mean zero and variance 0.2 and add it to both transformed potential outcomes. The common error cancels from the individual treatment

effect. The observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

We split this simulated population into sample ($D_i = 1$) and target ($D_i = 0$) populations using

$$P(D_i = 0 \mid X_i) = \text{logit}^{-1}(\vartheta \cdot (X_{1i} + X_{2i}))\,, \tag{42}$$

where $\vartheta \in \{0, 0.5, 1.0\}$ controls the degree of covariate shift between sample and target. At $\vartheta = 0$, population assignment is independent of the covariates, so the sample and target populations have the same covariate distribution. Their realized covariate distributions can still differ because of random sampling. When $\vartheta > 0$, the target covariate distribution places more mass on higher values of $X_{1i} + X_{2i}$ than the sample covariate distribution does. The design therefore creates continuous covariate shift, with ϑ controlling its magnitude.

We compare naive 2SLS, I2SLS, and DML using the estimation and prediction procedures described above. Both adjusted estimators use the same compliance weights.

For DML, we estimate the nuisance functions using XGBoost with two-fold cross-fitting. We divide the sample into two groups, train the nuisance models on one group, and predict the outcome, treatment, and instrument in the other, then reverse the roles of the groups. The resulting predictions are subtracted from the observed values to obtain the residuals used in estimation. We use 100 trees, a maximum tree depth of 3, and a learning rate of 0.1. These settings remain fixed across all parameter combinations.

For the compliance weights, the lower limit for X_1 is the larger of the two instrument-group-specific 0.5th percentiles in the sample, and the upper limit is the smaller of the two 99.5th percentiles. When a target observation falls outside this interval, we use the nearest boundary value only to evaluate the treatment probabilities that determine its compliance weight. We leave X_2 unchanged and use the observation's original covariates to evaluate its conditional effect and form the weighted target average. No target observations are dropped.

We repeat each parameter combination 300 times. In each replication, we evaluate prediction error relative to the true LATE among that replication's target compliers. Replication seeds are generated from a fixed seed of 42.

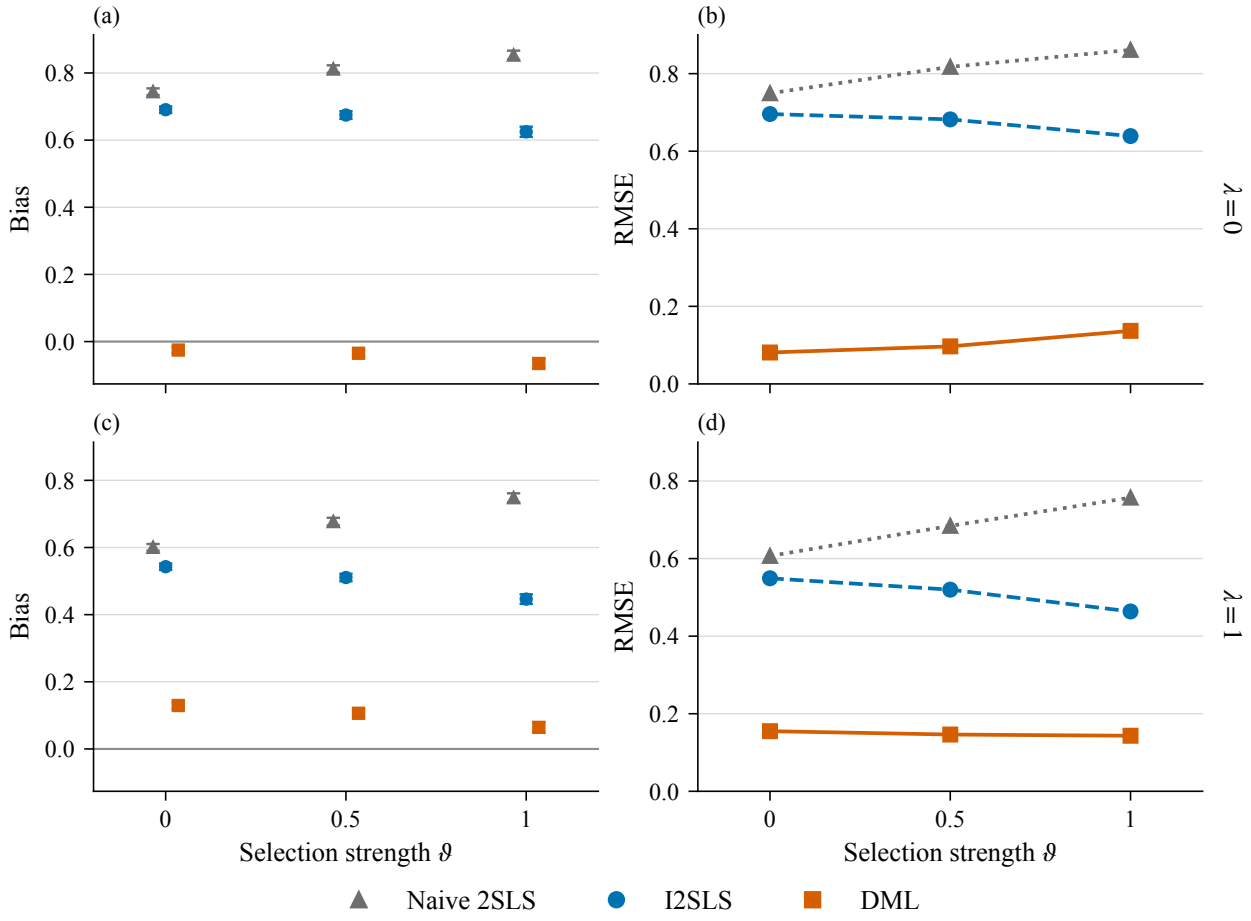


Figure 3: Bias and RMSE by Complier-Effect Setting and Selection Strength

Notes. 300 replications per combination, $N = 18{,}000$ per draw, with $\kappa = 1$ and $\gamma_y = 0$. Rows are the two complier-effect settings $\lambda = 0$ and $\lambda = 1$. The left column reports bias and the right column RMSE. The horizontal axis shows ϑ as defined in equation (42). Bias is the mean of the prediction error across replications, measured against that replication's own true target LATE. The bands in the bias column are plus or minus 1.96 Monte Carlo standard errors of the mean. The RMSE column has no bands. The design limitations described in Appendix I apply to this figure.

5.3 Baseline Results ($\kappa = 1, \lambda = 1, \gamma_y = 0$)

Figure 3 reports results for both the calibrated CATE ($\lambda = 1$) and its linear approximation ($\lambda = 0$), holding $\kappa = 1$ and $\gamma_y = 0$. We first focus on the calibrated specification. Equation (40) with $\lambda = 0$ defines the linear approximation, and Section 5.7 compares the two CATE settings.

Under the calibrated specification, DML has lower absolute bias and RMSE than both I2SLS and naive 2SLS at all three selection settings. DML nevertheless overpredicts the

target LATE, although its bias remains smaller than that of either 2SLS estimator.

As ϑ increases under the calibrated specification, the absolute bias of naive 2SLS rises, while that of I2SLS and DML falls. These declines in total absolute bias do not mean that larger sample–target differences reduce the errors involved in prediction. As Section 5.4 shows, several error components increase in magnitude as selection strengthens. Their contributions can nevertheless offset one another, producing a smaller total bias.

Appendix I evaluates whether the 95% confidence intervals for the predicted target LATE contain the true target LATE in approximately 95% of simulation replications. Appendix D.5 examines the sensitivity of the results to the compliance-weight model.

5.4 Bias Decomposition

Table 3: Prediction Error Components (Proposition 2): Card Calibration

ϑ		Term I	Term II	Term III	Term IV	Total bias
Panel A: $\lambda = 0$						
0	I2SLS	0.704 (0.004)	−0.001 (0.000)	0.000 (0.000)	−0.013 (0.002)	0.691 (0.005)
	DML	−0.004 (0.003)	0.000 (0.000)	0.000 (0.000)	−0.021 (0.002)	−0.025 (0.004)
0.5	I2SLS	0.690 (0.005)	0.014 (0.001)	0.000 (0.000)	−0.029 (0.002)	0.675 (0.006)
	DML	0.007 (0.004)	0.002 (0.001)	0.000 (0.000)	−0.044 (0.002)	−0.035 (0.005)
1	I2SLS	0.652 (0.006)	0.026 (0.002)	0.000 (0.000)	−0.053 (0.002)	0.625 (0.008)
	DML	0.006 (0.005)	0.004 (0.002)	0.000 (0.000)	−0.076 (0.003)	−0.065 (0.007)
Panel B: $\lambda = 1$						
0	I2SLS	0.555 (0.004)	−0.001 (0.000)	0.001 (0.001)	−0.012 (0.002)	0.543 (0.005)
	DML	0.148 (0.004)	−0.001 (0.000)	0.001 (0.001)	−0.019 (0.002)	0.129 (0.005)
0.5	I2SLS	0.546 (0.004)	−0.005 (0.001)	0.001 (0.001)	−0.031 (0.002)	0.511 (0.006)
	DML	0.150 (0.004)	−0.003 (0.001)	0.001 (0.001)	−0.041 (0.002)	0.106 (0.006)
1	I2SLS	0.518 (0.005)	−0.024 (0.001)	0.013 (0.002)	−0.060 (0.002)	0.446 (0.007)
	DML	0.146 (0.005)	−0.019 (0.002)	0.013 (0.002)	−0.075 (0.003)	0.064 (0.007)

Notes. Entries are means over 300 replications; parentheses report Monte Carlo standard errors. Terms I–IV are defined in Proposition 2 and sum to prediction error in each replication. Complier means and LATEs use the realized compliers in each replication. The reference coefficients are obtained from the compliance-weighted linear projection in the reference population. All settings hold $\kappa = 1$ and $\gamma_y = 0$. Rounding may prevent the displayed components from summing exactly to total bias.

Table 3 reports the four components of Proposition 2 for both CATE settings. We focus here on Panel B, which corresponds to the calibrated specification ($\lambda = 1$). Section 5.7

compares the two panels.

Under the calibrated specification, the main difference between I2SLS and DML arises in Term I, the error in estimating the LATE in the sample population before prediction is extended to the target population. Term I is larger for I2SLS than for DML at every selection setting. For DML, Term I remains approximately 0.15 across the three selection settings.

The decomposition also explains why total bias can fall as selection strengthens. For DML, Term I remains approximately unchanged, while Term IV becomes more negative, moving from -0.019 to -0.075. The growing negative contribution offsets the positive error in Term I, helping reduce total bias from 0.129 to 0.064. Thus, total bias falls even though compliance weighting error grows in magnitude. Both estimators use the same compliance weights, but their different estimated treatment-effect slopes produce different values of Term IV.

5.5 Instrument Propensity Nonlinearity (κ)

The next two comparisons hold the CATE linear ($\lambda = 0$) while varying nonlinearity in the instrument and outcome relationships. The $\lambda = 0$ case of equation (40) defines this linear CATE.

We examine how prediction performance changes as the instrument probability departs further from a linear relationship with the covariates. We vary $\kappa \in \{0, 1, 2\}$ in equation (39). The CATE remains linear ($\lambda = 0$), and the calibrated relationship between covariates and the outcome is unchanged ($\gamma_y = 0$). Thus, the differences across these settings arise from changes in the instrument-assignment process.

Figure 4 reports the resulting bias and RMSE. The I2SLS bias increases with κ at every selection setting, while DML bias changes much less. With no systematic selection ($\vartheta = 0$), for example, I2SLS bias rises from approximately zero at $\kappa = 0$ to 1.089 at $\kappa = 2$. Over the same settings, DML bias changes from -0.012 to -0.034. RMSE shows the same con-

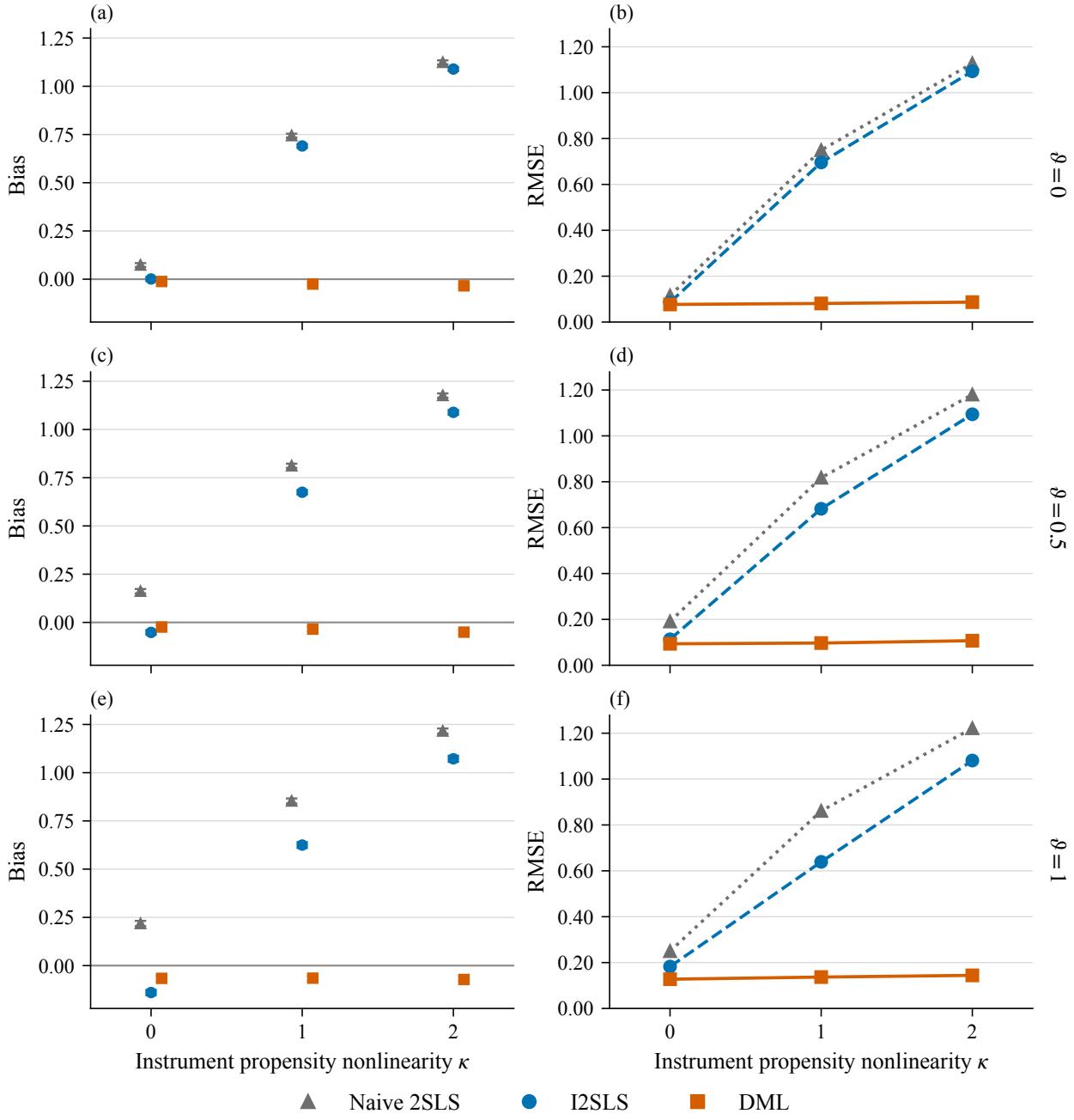


Figure 4: Bias and RMSE by Instrument Propensity Nonlinearity (κ) and Selection Strength (ϑ)

Notes. 300 replications per combination, $N = 18{,}000$ per draw, with $\lambda = 0$ and $\gamma_y = 0$. Rows are the three site-selection strengths ϑ. The left column reports bias and the right column RMSE. The horizontal axis shows $\kappa \in \{0, 1, 2\}$ as defined in equation (39). Bias is the mean of the prediction error across replications, measured against that replication's own true target LATE. The bands in the bias column are plus or minus 1.96 Monte Carlo standard errors of the mean. The design limitations described in Appendix I apply to this figure.

trast. The two adjusted estimators perform more similarly at $\kappa = 0$, and their performance separates as κ increases. Naive 2SLS also becomes more biased as κ increases.

These results are consistent with the role of instrument propensity misspecification described in Section 3.2. Linear covariate controls in I2SLS leave part of the nonlinear relationship between the instrument and covariates unaccounted for. This remaining component can affect the estimated treatment-effect coefficients. DML estimates the instrument probability flexibly, which helps explain its more stable performance in this comparison.

Changing the instrument probabilities also changes how frequently both instrument values occur among individuals with similar covariates and how strongly the instrument predicts treatment after accounting for covariates. The results therefore describe the combined effects of these changes as κ increases.

5.6 Outcome Nonlinearity (γ_y)

We next examine whether a nonlinear relationship between covariates and outcomes affects prediction even when the CATE remains linear. We vary $\gamma_y \in \{0, 2, 10\}$ using the common shift in equation (41). This shift preserves the individual treatment effects, compliance probabilities, and target LATE.

We hold the CATE linear ($\lambda = 0$) and use the instrument-propensity setting closest to a linear probability model among those considered ($\kappa = 0$). This comparison therefore focuses on changes in the outcome relationship while keeping the instrument probabilities and treatment effects fixed.

Figure 5 shows that increasing γ_y worsens prediction accuracy for I2SLS, while DML bias and RMSE remain nearly unchanged. For I2SLS, the loss of accuracy mainly reflects greater variability across replications when $\vartheta = 0$. With stronger selection, increasingly negative bias also contributes to its prediction error.

Linear covariate controls cannot capture the nonlinear term added to the outcome equation, whereas DML allows the conditional mean outcome to vary nonlinearly with covariates.

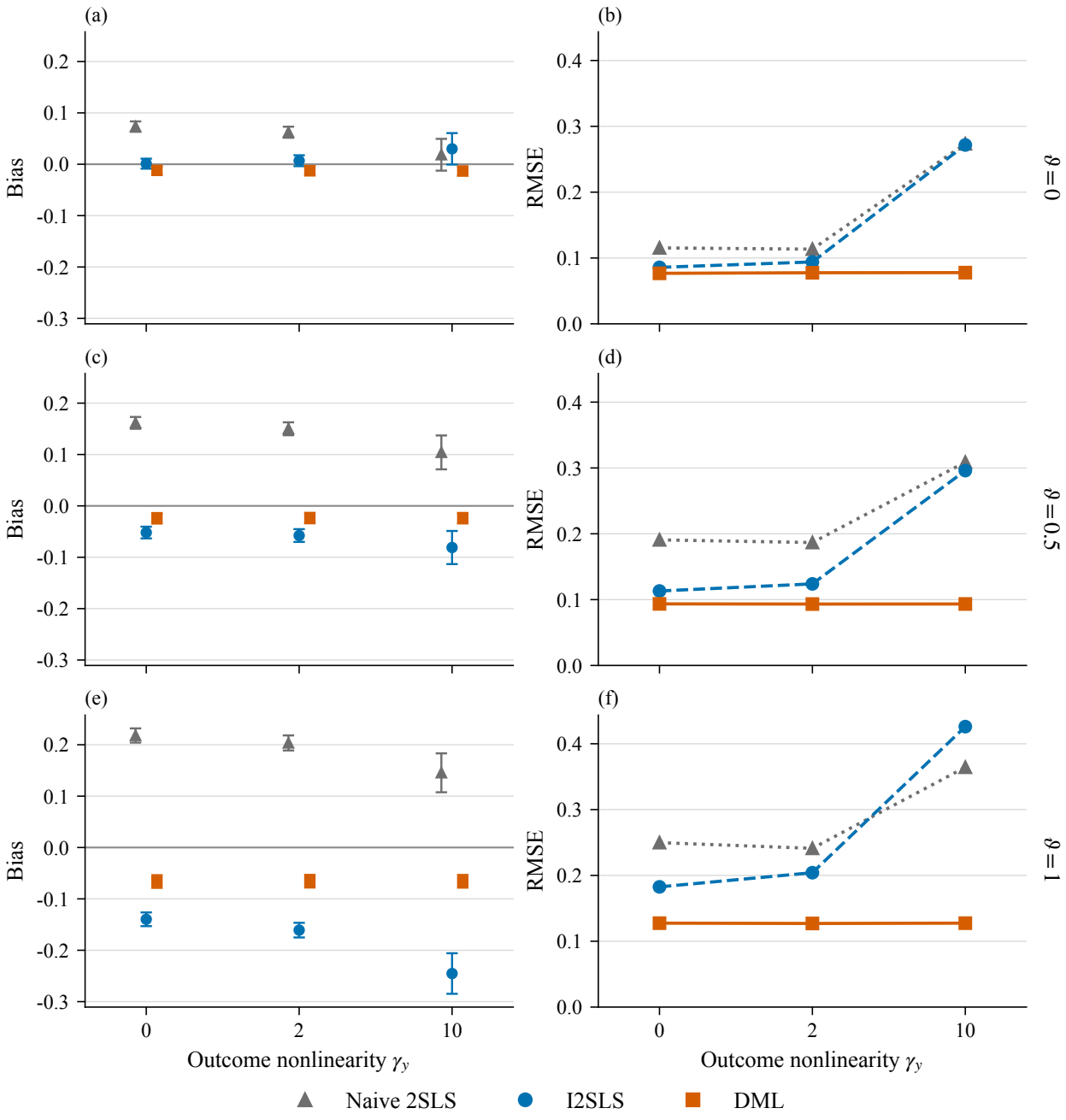


Figure 5: Bias and RMSE by Outcome Nonlinearity (γ_y) and Selection Strength (ϑ)

Notes. 300 replications per combination, $N = 18{,}000$ per draw, with $\kappa = 0$ and $\lambda = 0$. Rows are the three site-selection strengths ϑ. The left column reports bias and the right column RMSE. The horizontal axis shows $\gamma_y \in \{0, 2, 10\}$, placed at equal intervals and not drawn to scale. Bias is the mean of the prediction error across replications, measured against that replication's own true target LATE. The bands in the bias column are plus or minus 1.96 Monte Carlo standard errors of the mean. The design limitations described in Appendix I apply to this figure.

Equation (28) shows that the error in estimating the conditional mean outcome, $\widehat{q}(x) - q_0(x)$ with $q_0(x) = \mathbb{E}[Y \mid X = x]$, enters the DML moment multiplied by the error in estimating the instrument propensity. Accurate estimation of the instrument propensity therefore limits the effect of outcome-regression error, helping explain DML's stable performance in this comparison.

5.7 CATE Nonlinearity (λ)

We examine what happens when the conditional complier effect is nonlinear in the covariates. Both I2SLS and our baseline DML specification use a linear treatment-effect model. This comparison therefore examines a restriction shared by the two estimators. We compare $\lambda = 0$ and $\lambda = 1$ as defined in equation (40). The two settings have the same average effect among compliers in the reference population, 0.200. The target LATE can still differ between the two settings when selection changes the covariate distribution.

We now compare the $\lambda = 0$ and $\lambda = 1$ results in Figure 3 and Table 3, holding $\kappa = 1$ and $\gamma_y = 0$. Under the linear CATE, DML predictions are close to the target LATE, while both 2SLS estimators overpredict it. Most of the I2SLS bias comes from Term I, the error in estimating the LATE in the sample population. Although the CATE is linear, the instrument propensity retains its calibrated nonlinear form ($\kappa = 1$). The linear covariate controls in I2SLS do not fully capture this relationship, contributing to the error in estimating the LATE in the sample population. For DML, Term I is close to zero, and most of its remaining bias comes from compliance weighting error (Term IV). The magnitude of this weighting error increases as the sample and target covariate distributions diverge.

DML has higher RMSE under the calibrated nonlinear CATE than under the linear CATE at every selection setting. Its RMSE nevertheless remains lower than that of I2SLS. Comparing the two CATE settings for DML, Term I rises from approximately zero under the linear CATE to approximately 0.15 under the nonlinear CATE. Term III remains much smaller than Term I in this comparison, so the change in DML bias is concentrated in Term I.

This error is present even when there is no systematic difference between the sample and target covariate distributions.

When approximating the conditional effect, the linear OrthoIV model weights covariate values by both the complier probability and the instrument variance. Because LATE uses complier weights without this additional instrument-variance factor, the approximation errors need not average to zero among compliers when the CATE is nonlinear. This discrepancy can remain even when the nuisance functions are estimated accurately. Remark 1 in the appendix provides the formal explanation.

For I2SLS, both absolute bias and RMSE are lower under the nonlinear CATE than under the linear CATE. Table 3 shows, however, that the reduction in total bias does not reflect smaller contributions from every component. Term I falls, while Term III increases and Term IV does not consistently decline in magnitude. When selection is present, Term II also changes sign and offsets part of the positive contribution from Term I. The lower total bias therefore reflects both a reduction in Term I and changes in how the components offset one another.

We also examine whether allowing nonlinear treatment-effect models improves DML prediction, holding the estimated nuisance functions and compliance weights fixed. We compare the baseline linear model with two cubic specifications. The cubic models reduce estimated bias in some settings, but produce greater variability across replications and higher RMSE in all six settings. At the sample size considered here, allowing a more flexible treatment-effect model therefore does not improve overall prediction accuracy. Appendix H reports the specifications and full results.

5.8 Discussion

In the Card simulation, DML's advantage over I2SLS under the calibrated specification comes mainly from a smaller error in estimating the LATE in the sample population. The decomposition shows that much of the difference between the two estimators is already

present before prediction is extended to the target population. The separate comparisons also show that DML performance is more stable as the instrument probabilities change and as nonlinearity in the relationship between covariates and the outcome increases. Together, these results help explain why flexible nuisance models are useful in this calibration.

The comparison with the linear CATE clarifies a limitation of the DML specification used in the main results. Under the calibrated nonlinear CATE, the model produces error in estimating the LATE in the sample population even without systematic selection into the target population. Its linear treatment-effect specification cannot reproduce the nonlinear conditional effect, and the approximation errors need not average to zero among compliers. The cubic models allow nonlinear effects, but their prediction errors vary much more across replications. At the sample size considered here, this additional variability outweighs the reductions in estimated bias observed in some settings.

Compliance weighting helps explain the remaining patterns. The difference between the target LATE and the average conditional complier effect over the full target population shows that the choice of weights matters in this design. Using estimated compliance weights also contributes to prediction error. Under the calibrated nonlinear CATE, the negative compliance weighting error increasingly offsets the positive error in estimating the LATE in the sample population as selection strengthens. For DML, the error in estimating the LATE in the sample population remains approximately unchanged while total bias falls. The apparent improvement with stronger selection therefore reflects greater cancellation between errors. This is why the total bias alone gives an incomplete account of prediction performance in the Card simulation.

6 Conclusion

This paper studies prediction of a target population's LATE when only its covariates are observed. We state conditions under which conditional complier effects and compliance

probabilities from the sample population identify the target LATE. Our prediction-error decomposition distinguishes errors already present in the sample LATE from errors associated with differences between populations and with the estimated target-complier covariate mean. It provides a framework for understanding why a prediction succeeds or fails in the simulation designs.

The two simulation exercises show that the relative performance of DML and I2SLS depends on where functional-form misspecification enters the model. In the controlled Monte Carlo study, DML has lower RMSE than I2SLS in 26 of 30 settings with linear $g(X)$ and in 19 of 30 with step-function $g(X)$. In the Card simulation, DML performs more stably as the instrument probabilities and the outcome relationship change. However, flexible nuisance estimation does not remove the restrictions of a linear treatment-effect model. Allowing nonlinear treatment effects can reduce estimated bias, but the cubic specifications considered here increase variability and RMSE at the simulated sample size.

Successful prediction therefore requires more than accurate estimation of the sample LATE. It also requires a treatment-effect model that describes heterogeneity relevant to the target population and weights that reflect the composition of target compliers. The decomposition shows how errors in these components can reinforce or offset one another. A small total bias can arise from cancellation, so examining its components gives information that an overall performance measure alone cannot provide.

Future work can develop methods for choosing treatment-effect models without observing the target LATE and for inference that accounts jointly for coefficient estimation, compliance-weight estimation, and the observed target covariate sample. Sensitivity analysis can also assess prediction when conditional complier effects differ across populations or compliance probabilities are not proportional. These extensions would connect the simulation findings to prediction in applications where the target effect is unknown.

References

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

Angrist, Joshua D. and Ivan Fernandez-Val, "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in Daron Acemoglu, Manuel Arellano, and Eddie Dekel, eds., *Advances in Economics and Econometrics: Tenth World Congress*, Vol. 3, Cambridge: Cambridge University Press, 2013, pp. 401–434.

_ **and Jörn-Steffen Pischke**, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press, 2009.

Balietti, Anca and Souvik Datta, "The Impact of Indoor Solid Fuel Use on the Stunting of Indian Children," 2017. Unpublished manuscript, Center for International Development, Harvard Kennedy School of Government.

Battocchi, Keith, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," https://github.com/py-why/EconML 2019. Version 0.16.0.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, "When Is TSLS Actually LATE?," *The Review of Economic Studies*, May 2026, p. rdag029.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, 1995, pp. 201–222.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 2018, *21* (1), C1–C68.

Frölich, Markus, "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 2007, *139* (1), 35–75. Endogeneity, instruments and identification.

Galiani, Sebastian and Brian Quistorff, "Assessing External Validity in Practice," *Research in Economics*, September 2024, *78* (3), 100964.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer, "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, 2005, *125* (1), 241–270.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Kwon, Alexander and Kyungtae Lee, "External Validity in an Instrumental Variable Setting," *Evaluation Review*, 2025, *49* (6), 1000–1020.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Rudolph, Kara E. and Mark J. van der Laan, "Robust Estimation of Encouragement Design Intervention Effects Transported Across Sites," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 11 2017, *79* (5), 1509–1525.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586.

Su, Qinghe and Mehtabul Azam, "Does Access to Liquefied Petroleum Gas (LPG) Reduce the Household Burden of Women? Evidence from India," *Energy Economics*, 2023, *119*, 106529.

Syrgkanis, Vasilis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis, "Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments," in "Advances in Neural Information Processing Systems," Vol. 32 2019.

Zhao, Anqi, Peng Ding, and Fan Li, "Interacted Two-Stage Least Squares with Treatment Effect Heterogeneity," 2025.

A Magnitude of the Nonlinearity Term

Define the nonlinear residual $r(x) = \tau_S(x) - \alpha_0 - \alpha^\top x$. When the conditional effect is linear and the reference coefficients describe that effect, $r(x) = 0$ and Term III vanishes. For a nonlinear effect, the contribution of $r(x)$ depends on how the complier covariate distribution changes. When F_T is absolutely continuous with respect to F_S, let $\rho(x) = dF_T/dF_S(x)$ be the density ratio. Then $\mathbb{E}_{F_S}[\rho(X)] = \int dF_T = 1$. Suppose $r(X)$ and $\rho(X)$ are square integrable under F_S. Using the density ratio to express both integrals under F_S gives

$$\begin{aligned}\text{Term III} &= -\int r(x)\, d(F_T - F_S)(x) \\ &= -\mathbb{E}_{F_S}[r(X)\rho(X)] + \mathbb{E}_{F_S}[r(X)] \\ &= -\mathrm{Cov}_{F_S}(r(X), \rho(X)).\end{aligned}$$

By Cauchy–Schwarz,

$$|\text{Term III}| \le \sqrt{\mathrm{Var}_{F_S}(r(X))} \cdot \sqrt{\mathrm{Var}_{F_S}(\rho(X))}.$$

The first factor measures the severity of nonlinearity in τ_S. The second measures the magnitude of the distribution shift between sample and target compliers. With the reference coefficients chosen to recover a linear conditional effect, nonlinearity is necessary but not sufficient for a nonzero Term III. Term III vanishes when $F_T = F_S$, so that $\rho \equiv 1$, and whenever r and ρ are uncorrelated under F_S.

B Linear OrthoIV Population Projection

This appendix explains why a linear OrthoIV treatment-effect model can misestimate LATE when the CATE is nonlinear. We derive the equations determining its coefficients using the true conditional means in the sample population. Write q_0, p_0 and h_0 for the sample-population conditional mean functions of equations (18) to (20). Because Z_i is binary, any

integrable variable A_i satisfies the following identity where $0 < h_0(x) < 1$:

$$\begin{aligned}\mathrm{Cov}(A_i, Z_i \mid X_i = x, D_i = 1) &= h_0(x)\{1 - h_0(x)\} \\ &\quad \times \big\{\mathbb{E}[A_i \mid Z_i = 1, X_i = x, D_i = 1] - \mathbb{E}[A_i \mid Z_i = 0, X_i = x, D_i = 1]\big\}.\end{aligned}$$

Under Assumption 1 this gives the two conditional covariances

$$\mathrm{Cov}(Y_i, Z_i \mid X_i, D_i = 1) = h_0(X_i)\{1 - h_0(X_i)\}\pi_S(X_i)\tau_S(X_i),$$

$$\mathrm{Cov}(T_i, Z_i \mid X_i, D_i = 1) = h_0(X_i)\{1 - h_0(X_i)\}\pi_S(X_i).$$

The conditional Wald ratio therefore equals $\tau_S(x)$ where $0 < h_0(x) < 1$ and $\pi_S(x) > 0$. The denominator is zero outside those values, so the ratio is not defined there. The two factors have distinct roles. The term $h_0(x)\{1 - h_0(x)\}$ measures conditional instrument variation, and $\pi_S(x)$ measures the conditional treatment response to the instrument. Covariate values with no instrument variation or no sample compliers receive zero weight in the weighted normal equations in the sample population. This statement concerns identification in the sample population and does not by itself establish transport to the target population.

Let $\phi(X_i)$ contain an intercept and the covariates or basis terms in the treatment-effect model, so that $\phi(X_i) = (1, X_i^\top)^\top$ for the linear specification. With the `projection=False` option, the OrthoIV moment condition is

$$\mathbb{E}_S\big[\phi(X_i)\{Z_i - h_0(X_i)\}\{Y_i - q_0(X_i) - (T_i - p_0(X_i))\phi(X_i)^\top\beta^\dagger\}\big] = 0. \tag{43}$$

The residual instrument has conditional mean zero, so the terms in q_0 and p_0 drop out after conditioning on X_i. The moment condition can then be written as

$$\mathbb{E}_S\big[\phi(X_i)\left\{\mathrm{Cov}(Y_i, Z_i \mid X_i, D_i = 1) - \phi(X_i)^\top\beta^\dagger\mathrm{Cov}(T_i, Z_i \mid X_i, D_i = 1)\right\}\big] = 0. \tag{44}$$

Let $Q = \mathbb{E}_S[h_0(1-h_0)\pi_S\,\phi\phi^\top]$ be the weighted regressor second-moment matrix.

Lemma 1 (Weighted normal equations)**.** *Suppose the instrument is binary and Assumption 1 holds in the sample population, the treatment-effect coefficients solve the* `projection=False` *moment condition (43), the matrix Q and the corresponding right-hand-side moments are finite, and Q has full rank. Then the population OrthoIV treatment-effect function $\tau^\dagger(x) = \phi(x)^\top\beta^\dagger$ satisfies*

$$\mathbb{E}_S[h_0(1-h_0)\pi_S\,\phi\{\tau_S - \phi^\top\beta^\dagger\}] = 0.$$

Proof. Substitute the two conditional covariances into the covariance form (44). Factoring $\beta^\dagger$ out of the resulting weighted normal equations gives $Q\beta^\dagger$ on the coefficient side, and the full rank of Q identifies the coefficient vector. □

The full-rank condition on Q is a regressor-variation and no-collinearity condition. It is not implied by IV validity. Take an intercept and a scalar X. Positive total weight and positive weighted variance of X are then necessary and sufficient for full rank.

Remark 1 (Weighted linear approximation of the complier effect)**.** *With the true nuisance functions, our baseline OrthoIV specification selects a weighted linear approximation to $\tau_S(x)$, as shown in Lemma 1. The weights are $h_0(x)\{1-h_0(x)\}\pi_S(x)$. They give greater weight to covariate values with more instrument variation and a larger complier share. Compared with the compliance weights used to calculate the target LATE, these weights include the additional instrument-variation factor $h_0(x)\{1-h_0(x)\}$.*

Write the population linear approximation as $\tau^\dagger(x) = \phi(x)^\top\beta^\dagger$ and its error as $r_{\mathrm{DML}}(x) = \tau_S(x) - \tau^\dagger(x)$. When $\tau_S(x)$ is linear, this approximation recovers the conditional-effect function on the identified support. When $\tau_S(x)$ is nonlinear, the approximation error is orthogonal to the included regressors under the weights $h_0(1-h_0)\pi_S$. Positive and negative approximation errors can therefore cancel under these weights, but their average need not be zero under the covariate distribution of target compliers. Consequently, averaging the linear approximation over target compliers need not recover the target LATE.

This error can remain even when the sample and target populations have identical covariate distributions. Flexible nuisance estimation alone does not remove it.

C Cooking Fuel Application

C.1 Data and Instrument

We evaluate external validity using the effect of solid fuel use on average daily cooking time (measured in hours per day) across six developing countries: Ethiopia (n = 3,392), Honduras (n = 2,189), Kenya (n = 3,417), Cambodia (n = 3,100), Nepal (n = 5,602), and Zambia (n = 2,793). The data come from the World Bank's Multi-Tier Framework (MTF) survey, which provides nationally representative samples with harmonized variables across countries. The treatment T_i is a binary indicator of solid fuel use for cooking, equal to 1 if household i uses solid fuel and 0 otherwise. The outcome Y_i is the household's average daily cooking time in hours, converted from the minutes recorded in the survey. The measure aggregates cooking time across the household's members, and individual-level time allocation is not recorded. The covariate vector X_i contains eleven household characteristics: the mean age of women in the household, an indicator for a male household head, an indicator for post-secondary education taken as the household maximum, the number of children, household size, the share of women in the household, an urban indicator, an indicator for a stove that is not clean, house structure, the aggregate cooking-solution tier, and a wealth index. The sample excludes households with a missing value on the outcome, the treatment, the instrument, or any of these covariates.

The instrumental variable measures neighborhood-level fuel availability using a leave-one-out spatial average: for each household i in country s, the IV is the mean fuel usage rate of other households in the same neighborhood:

$$Z_{is} = \frac{1}{|J| - 1} \sum_{j \in J,\, j \neq i} T_j, \tag{45}$$

where J is the set of households in i's neighborhood. For regression, we construct a binary instrument $Z_{is}^d = \mathbf{1}\{Z_{is} > \bar{Z}_s\}$. The exclusion restriction requires that neighborhood fuel availability affects cooking time only through a household's own fuel choice. The leave-one-out construction mitigates mechanical correlation by excluding household i's own usage from the average. A causal interpretation requires that, conditional on the household characteristics in X_i, the remaining variation in Z_{is} affect cooking time only through fuel adoption. The main threat to exclusion is that neighborhood fuel prevalence correlates with local infrastructure or economic conditions that directly affect cooking time. The wealth index and the urban indicator in X_i are intended to absorb these confounders. Prior studies using this instrument class include Balietti and Datta (2017) for solid fuels and Su and Azam (2023) for LPG. We assess the instrument's association with treatment, controlling for covariates, using the first-stage F-statistic. It tests whether the instrument coefficient is zero in a regression of T_i on Z_{is}^d and X_i. The country-specific statistics, calculated using heteroskedasticity-robust standard errors, are Ethiopia ($F = 41.8$), Honduras ($F = 70.3$), Kenya ($F = 75.9$), Cambodia ($F = 104.7$), Nepal ($F = 9.4$), and Zambia ($F = 35.5$). Nepal's statistic is 9.4, while the other five values range from 35.5 to 104.7. This contrast warrants particular caution about weak-instrument bias and inference (Staiger and Stock, 1997). Nepal's smaller F-statistic is consistent with the imprecision of Nepal's estimated target LATE (SE = 6.68). The external validity exercise is a methods comparison, and the instrument is applied consistently across all six countries. All three estimators rely on the same underlying instrument validity conditions, so a serious violation of the exclusion restriction would undermine the causal interpretation of the estimated target LATE and of every prediction alike. We do not claim that the resulting quantitative bias would be equal across estimators, since they residualize, weight, and restrict functional form differently.

We predict each of the six countries in turn using a leave-one-country-out design. For target country c, the target is country c itself and the $D_i = 1$ sample pools the five remaining countries. Each country therefore serves once as the target and five times in the sample.

Covariates enter demeaned, and the prediction moves the CATE from the sample covariate mean to country c's covariate mean.

We estimate I2SLS on the pooled five-country sample. The specification adds country fixed effects and interactions of the country indicators with the covariates demeaned at the sample mean, $\ddot{X}_i = X_i - \bar{X}_S$, to equation (15), so that level differences across the five sample countries are absorbed. T_i and $T_i\ddot{X}_i$ are instrumented by Z^d_{is} and $Z^d_{is}\ddot{X}_i$. The prediction uses the estimated linear treatment-effect model at country c's covariate mean, and its standard error is obtained by the delta method applied to that prediction.

We estimate DML on the same pooled five-country sample, with X_i entering untransformed and the nuisance functions estimated using gradient-boosted trees. The prediction averages the estimated CATE $\hat{\tau}_{\text{DML}}(X_i)$ over the households of country c. It therefore integrates against the empirical target covariate distribution. Its standard error is obtained from a bootstrap with 100 replications. The target LATE is estimated separately by 2SLS on country c alone, with linear controls for X_i and T_i instrumented by Z^d_{is}. Sections 4 and 5 compare predictions against the true target LATE computed from the DGP. The case study has no known target truth, and the separate 2SLS estimate is used for comparison. Table 4 reports its standard errors.

Both predictions reported in this appendix use the target population covariate mean to calculate the predicted effect. These estimates were re-estimated for this paper using the data construction and leave-one-country-out design described in Section 1.1. Both predictions average over the full target population, without the compliance weights of Section 2.3. The unweighted and compliance-weighted covariate means coincide when compliance does not vary with covariates in the target country.

The DML hyperparameters are not held fixed across countries. For each target country, we adjusted them to bring the predicted LATE as close as possible to that country's estimated target LATE. We estimate that target LATE from country c's own data. A practitioner predicting an unstudied target would not observe it. The DML model is estimated

using the other five countries, so it produces a prediction for a held-out target country. Its hyperparameter selection nevertheless uses information about the target effect. We call this row target-tuned DML and interpret it as a favorable descriptive comparison. It shows the performance of the candidate specifications under target-informed selection, rather than an independent evaluation for an unstudied target.

C.2 Prediction Results

Table 4: The Effect of Solid Fuel on Cooking Time: Predictions and Estimated Target LATEs

	Ethiopia	Honduras	Kenya	Cambodia	Nepal	Zambia
Panel A: Predictions and Estimated Target LATEs						
Estimated target LATE	1.652	2.009	1.450	0.628	15.438	3.228
	(0.469)	(0.634)	(0.532)	(0.151)	(6.684)	(0.731)
Predicted (I2SLS)	2.259	3.756	2.594	2.397	2.646	2.919
	(0.736)	(0.532)	(0.634)	(0.752)	(1.010)	(0.677)
DML (target-tuned)	1.628	2.005	1.405	1.115	1.948	2.119
	(0.735)	(1.184)	(0.444)	(0.151)	(1.097)	(1.248)
Panel B: Differences Relative to the Estimated Target LATE						
I2SLS: % Diff	36.7%	86.9%	78.9%	281.7%	−82.9%	−9.6%
Std. diff.	0.696	2.111	1.382	2.306	−1.892	−0.310
DML: % Diff	−1.5%	−0.2%	−3.1%	77.5%	−87.4%	−34.4%
Std. diff.	−0.028	−0.003	−0.065	2.281	−1.992	−0.767

Notes. All estimates were recomputed for this paper. Standard errors are in parentheses. The estimated target LATE is a separate 2SLS estimate with linear controls and heteroskedasticity-robust standard errors. Predicted standard errors use a 100-replication bootstrap for DML and the delta method for I2SLS. "% Diff" is the prediction minus the estimated target LATE, divided by the estimated target LATE. "Std. diff." divides the same difference by $\sqrt{\mathrm{SE}^2_{\mathrm{pred}} + \mathrm{SE}^2_{\mathrm{target}}}$ under an independence approximation and is descriptive, not a test statistic. The DML model is estimated on the other five countries, but its hyperparameters are selected using the estimated target LATE. The row is therefore a target-informed comparison, not an independent evaluation of prediction performance. Its standard error does not account for that tuning. I2SLS uses no target outcome, treatment or instrument data. Both predictions use the target population covariate mean rather than the target covariate mean calculated using the compliance weights of Section 2.3.

Table 4 reports the predictions and their differences from each country's estimated target

LATE. Figure 1 displays the same estimates. The target-tuned DML prediction is closer to the estimated target LATE than I2SLS in Ethiopia, Honduras, Kenya and Cambodia. It is less close in Nepal and Zambia. Thus, the favorable tuning exercise improves agreement in some countries but does not eliminate uneven performance across targets.

Nepal has the largest gap between both predictions and the estimated target LATE. Its estimated target LATE is 15.44 hours per day, compared with predictions of 1.95 for DML and 2.65 for I2SLS. This estimate is also the least precise of the six, with a standard error of 6.68, and Nepal has the weakest instrument–treatment association. These features make uncertainty in the estimated target LATE important when interpreting the prediction gap. The observed outcome range alone does not establish that the causal effect estimate is invalid.

Several explanations could contribute to this gap. The estimated target LATE may be affected by weak-instrument bias or sampling variability. Conditional complier effects may differ across countries even at the same observed covariate values. Target covariate values may also be poorly represented in the observed sample. A lack of observations in a realized sample needs to be distinguished from a failure of the population support condition in Assumption 4.

The use of full-population averages is another possible contributor. When compliance probabilities and conditional complier effects are correlated across target covariate values, averaging over all target households differs from averaging over target compliers. Section 2.3 explains this distinction and the compliance weights required under the transport assumptions. The results here do not separate these explanations. They motivate the controlled simulations, where the true target LATE and the components of prediction error are known.

D Compliance Weights in the Simulation Designs

This appendix reports the aggregation gap G defined in Section 2.3 for the Monte Carlo and Card simulation designs. It also reports weight and support diagnostics and a comparison of covariate limits for the Monte Carlo design, and the sensitivity of the Card results to the compliance-weight model.

D.1 Aggregation Gap in the Monte Carlo Design

Table 5: Target LATE, Target-Population Mean Conditional Effect, and the Aggregation Gap

Scale	ϑ	τ_T	$\mathbb{E}_T[\tau_S(X_i)]$	G	SE(G)
1	0	7.461	6.997	0.464	0.004
1	0.25	8.676	8.232	0.444	0.006
1	0.5	9.769	9.356	0.413	0.004
1	0.75	10.702	10.327	0.375	0.004
1	1	11.484	11.131	0.353	0.002
3	0	11.23	7.01	4.22	0.01
3	0.25	20.03	17.00	3.03	0.01
3	0.5	25.00	22.91	2.09	0.01
3	0.75	27.55	25.99	1.56	0.01
3	1	28.92	27.67	1.24	0.02
10	0	49.31	7.06	42.25	0.05
10	0.25	82.00	72.57	9.42	0.03
10	0.5	86.23	82.16	4.07	0.03
10	0.75	87.36	84.59	2.78	0.02
10	1	87.86	85.53	2.33	0.02

Notes. τ_T is the target LATE, $\mathbb{E}_T[\tau_S(X_i)]$ is the unweighted average of conditional complier effects over the target population, and $G = \tau_T - \mathbb{E}_T[\tau_S(X_i)]$ is the aggregation gap defined in Section 2.3. Entries are population constants averaged over eight draws of 2,000,000 units. SE(G) is the standard error of that mean, not a Monte Carlo standard error over estimation replications. Decimals follow the largest standard error among the three means in each row.

Table 5 compares the true LATE in the target population, the average treatment effect among target compliers, with the average of the true conditional complier effect $\tau_S(X)$ over the entire target population. The conditional complier effect is the same in both populations

by construction, but the two averages use different covariate distributions. Their values change with the covariate scale σ_X and selection strength ϑ. Because complier probabilities vary with X, averaging over all target observations gives a different result from averaging over target compliers. The gap G shows why compliance weighting matters in this design.

D.2 Aggregation Gap in the Card-Calibrated Design

Table 6 shows why compliance weighting matters in this design. When $\lambda = 0$, averaging conditional complier effects over the whole target population understates the target LATE: the gap G ranges from 0.122 to 0.129 across the three site-selection settings. When $\lambda = 1$, the same unweighted average overstates the target LATE, with G between -0.084 and -0.083. These differences are substantial relative to the target LATE. We calculate them directly from the known simulation model.

Table 6: Target LATE, Target-Population Mean Conditional Effect, and the Aggregation Gap

λ	ϑ	τ_T	$\mathbb{E}_T[\tau_S(X_i)]$	G
0	0	0.200	0.078	0.122
0	0.5	0.173	0.047	0.126
0	1	0.162	0.032	0.129
1	0	0.200	0.283	−0.083
1	0.5	0.172	0.256	−0.084
1	1	0.158	0.241	−0.083

Notes. Entries are population constants on the 177-cell support and have no Monte Carlo standard error. τ_T is the target LATE, $\mathbb{E}_T[\tau_S(X_i)]$ is the unweighted average of conditional complier effects over the target population, and $G = \tau_T - \mathbb{E}_T[\tau_S(X_i)]$ is the aggregation gap defined in Section 2.3. λ varies the complier effect and ϑ varies site selection, while $\kappa = 1$ and $\gamma_y = 0$.

Results are conditional on the chosen calibration. Appendix I states the design limits that apply to this table and to Tables 3 and 13.

D.3 Weight and Support Diagnostics in the Monte Carlo Design

This subsection examines how the compliance-weight procedure changes estimated differences in treatment probabilities at target covariate values. Table 7 reports the same sixteen combinations highlighted in the main Monte Carlo results.

Raw differences are evaluated at the original target covariates. Boundary-adjusted differences evaluate the same models at the nearest permitted value when a target covariate falls outside the chosen interval. Any remaining negative differences are then set to zero. The original target covariates are retained when calculating the weighted target mean.

The negative share measures how often the estimated difference is below zero. The negative-to-positive mass ratio measures the size of those negative values relative to the positive values. The outside share measures how often boundary adjustment is used. The ESS column describes concentration before boundary adjustment. For this diagnostic, define $w_i^{\text{raw}} = \max\{\widehat{m}_1(X_i) - \widehat{m}_0(X_i), 0\}$ using the original target covariates. When these weights have a positive sum, their effective sample size is

$$\text{ESS} = \frac{(\sum_{i:D_i=0} w_i^{\text{raw}})^2}{\sum_{i:D_i=0} (w_i^{\text{raw}})^2}.$$

We report ESS as a percentage of the target sample size. Equal positive weights give 100%; concentration on fewer observations lowers this percentage. This column describes the positive part of the raw differences, before the boundary adjustment used to construct the final prediction weights.

In all settings at scale 1, and at scale 10 with $\vartheta = 0$, the raw negative share is exactly zero, so no target observation has a negative raw difference. At scale 10 with $\vartheta = 1$, about 21–23% of target observations have a negative raw difference, and the negative-to-positive mass ratio is about 4.2–4.6%. Boundary adjustment reduces the negative share in these settings to about 0.02% or less, but this reduction alone does not show that the resulting weights accurately approximate complier probabilities. The adjustment is used much more frequently in some high-scale settings. For example, at scale 10 with the cubic instrument specification, about 70–79% of target observations are outside the chosen interval, compared with about 1–5% at scale 1. The large outside shares motivate the comparison of alternative covariate limits below.

Table 7: Compliance-weight and support diagnostics

				Raw compliance			Boundary-adjusted		
Scale	Instrument	Outcome	ϑ	Negative share (%)	Neg./pos. mass (%)	ESS (%)	Negative share (%)	Zero share (%)	Outside share (%)
1	linear	linear	0	0.000	0.0000	97.874	0.0000	0.0000	1.046
1	linear	linear	1	0.000	0.0000	97.253	0.0000	0.0000	5.040
1	linear	step	0	0.000	0.0000	97.875	0.0000	0.0000	1.044
1	linear	step	1	0.000	0.0000	97.263	0.0000	0.0000	5.007
1	cubic	linear	0	0.000	0.0000	97.902	0.0000	0.0000	1.050
1	cubic	linear	1	0.000	0.0000	97.467	0.0000	0.0000	4.998
1	cubic	step	0	0.000	0.0000	97.812	0.0000	0.0000	1.036
1	cubic	step	1	0.000	0.0000	97.424	0.0000	0.0000	4.989
10	linear	linear	0	0.000	0.0000	91.536	0.0000	0.0000	13.318
10	linear	linear	1	23.238	4.4978	70.307	0.0000	0.0000	78.616
10	linear	step	0	0.000	0.0000	91.475	0.0000	0.0000	13.397
10	linear	step	1	22.010	4.1621	70.896	0.0000	0.0000	78.693
10	cubic	linear	0	0.000	0.0000	92.813	0.0000	0.0000	70.034
10	cubic	linear	1	21.003	4.2638	73.592	0.0198	0.0197	78.763
10	cubic	step	0	0.000	0.0000	92.864	0.0000	0.0000	69.996
10	cubic	step	1	21.147	4.5988	73.142	0.0211	0.0211	78.793

Notes: The raw and boundary-adjusted negative shares pool target observations across replications. The negative-to-positive mass ratio is the sum of absolute negative values divided by the sum of positive values before boundary adjustment. For ESS, negative raw differences are set to zero before any boundary adjustment. We calculate ESS as a fraction of the target sample size within each replication, average these fractions, and report them as percentages. The zero share is the mean replication-level share of zero weights after setting remaining negative boundary-adjusted values to zero. The outside share is the mean replication-level share whose original target covariate lies outside the intersection interval and is therefore evaluated at its boundary. These metrics describe the behavior of the estimated and adjusted weights and do not establish proportional compliance, support identification or prediction accuracy.

D.4 Comparison of Covariate Limits for Compliance Weights

We compare three choices of covariate limits. The estimated treatment probabilities and simulated samples are the same in each comparison. Edge uses the 0.5th and 99.5th percentiles of the covariate observations from the sample population. Overlap spans the sample covariate values whose estimated instrument probability lies between 0.01 and 0.99, as defined in Section 4.1. Intersection uses the range satisfying both rules. Each rule changes where the same probability models are evaluated for target observations outside its interval.

Table 8 shows that these choices can materially change prediction performance. At scale 1, Edge and Intersection give the same reported results. At scale 10 with the cubic instrument specification and no systematic selection, Intersection matches Overlap and produces more negative bias and higher RMSE than Edge for both estimators. The relative performance of

the intervals differs across other settings. The comparison therefore shows that the boundary rule is part of the estimator and can affect its results.

Table 8: Sensitivity of predictions to the evaluation interval

Scale	Instrument	Outcome	ϑ	Method	Bias			RMSE		
					Edge	Overlap	Intersection	Edge	Overlap	Intersection
1	linear	linear	0	DML	0.164	0.143	0.164	0.513	0.508	0.513
1	linear	linear	0	I2SLS	0.100	0.080	0.100	0.473	0.470	0.473
1	linear	linear	1	DML	−0.601	−0.762	−0.601	0.808	0.933	0.808
1	linear	linear	1	I2SLS	−0.788	−0.946	−0.788	0.962	1.094	0.962
1	linear	step	0	DML	0.151	0.130	0.151	0.512	0.507	0.512
1	linear	step	0	I2SLS	0.109	0.088	0.109	0.485	0.481	0.485
1	linear	step	1	DML	−0.618	−0.779	−0.618	0.829	0.953	0.829
1	linear	step	1	I2SLS	−0.040	−0.212	−0.040	0.585	0.618	0.585
1	cubic	linear	0	DML	0.186	0.166	0.186	0.520	0.513	0.520
1	cubic	linear	0	I2SLS	0.156	0.136	0.156	0.484	0.478	0.484
1	cubic	linear	1	DML	−0.590	−0.745	−0.590	0.797	0.917	0.797
1	cubic	linear	1	I2SLS	−0.868	−1.017	−0.868	1.037	1.164	1.037
1	cubic	step	0	DML	0.215	0.195	0.215	0.544	0.537	0.544
1	cubic	step	0	I2SLS	−0.016	−0.037	−0.016	0.488	0.489	0.488
1	cubic	step	1	DML	−0.574	−0.730	−0.574	0.791	0.909	0.791
1	cubic	step	1	I2SLS	0.514	0.329	0.514	0.786	0.675	0.786
10	linear	linear	0	DML	−16.505	−14.148	−14.142	17.385	15.126	15.121
10	linear	linear	0	I2SLS	−17.152	−14.798	−14.792	17.940	15.664	15.658
10	linear	linear	1	DML	1.089	−17.314	1.089	9.104	34.063	9.104
10	linear	linear	1	I2SLS	−5.312	−24.349	−5.312	84.736	57.827	84.736
10	linear	step	0	DML	−16.038	−13.731	−13.726	16.831	14.647	14.641
10	linear	step	0	I2SLS	−13.020	−10.771	−10.766	13.997	11.930	11.925
10	linear	step	1	DML	−0.269	−17.588	−0.269	4.965	31.826	4.965
10	linear	step	1	I2SLS	65.295	28.411	65.295	742.805	592.880	742.805
10	cubic	linear	0	DML	−19.560	−25.918	−25.918	20.579	26.461	26.461
10	cubic	linear	0	I2SLS	−20.764	−27.125	−27.125	21.640	27.588	27.588
10	cubic	linear	1	DML	1.498	0.461	1.489	5.805	9.317	5.801
10	cubic	linear	1	I2SLS	28.760	26.632	28.747	747.693	727.178	747.619
10	cubic	step	0	DML	−19.704	−26.422	−26.422	20.680	26.931	26.931
10	cubic	step	0	I2SLS	−10.038	−16.769	−16.769	12.065	17.576	17.576
10	cubic	step	1	DML	−0.187	−1.270	−0.198	5.839	9.569	5.837
10	cubic	step	1	I2SLS	95.243	92.103	95.143	12625.784	12696.660	12624.697

Notes: Bias is mean prediction error and RMSE is root mean squared prediction error over 500 paired replications. Edge uses the 0.5th and 99.5th percentiles of the covariate observations drawn from the sample population, Overlap uses the instrument-overlap interval, and Intersection uses their intersection. The comparison isolates the evaluation interval and does not establish that one interval is generally preferred.

D.5 Sensitivity to the Compliance-Weight Model in the Card Design

This subsection examines whether the choice of compliance-weight model changes the comparison between DML and I2SLS. Table 9 replaces the baseline scaled-logit model with a spline-logit model using the same simulation replications. Within each weight specification, both estimators use the same compliance weights.

The ranking is unchanged in all six settings. DML has lower RMSE than I2SLS under both weight models. The magnitude of prediction error nevertheless depends on the weight specification. For I2SLS, spline-logit weights increase both bias and RMSE in every setting. For DML, they slightly reduce or leave reported RMSE unchanged when the CATE is linear, but increase RMSE when the CATE is nonlinear. For example, with a nonlinear CATE and the strongest selection, DML RMSE rises from 0.143 to 0.202.

The comparison shows that the estimator ranking in these settings is stable to this alternative weight model. It also shows that changing how compliance is estimated can affect target prediction. Prediction performance alone does not establish which model estimates compliance probabilities more accurately.

E Additional Monte Carlo Details

E.1 The Nonlinearity Term in the Monte Carlo DGP

In the Monte Carlo design of Section 4.1, the individual effect satisfies equation (33), so the complier CATE is $\tau^*(x) = 10x + 7$, linear in x regardless of the outcome specification. The compliance-type offsets are constants within the complier population and therefore do not curve τ^*. Therefore $r(x) \equiv 0$ and the population CATE nonlinearity $\times$ distribution shift component vanishes for all simulation combinations. The step function under $y_{\text{spec}} = \text{step}$ enters the outcome level via $g(X_i) = 5 \cdot \mathbf{1}\{X_i > 0\}$ and cancels in the treatment-effect

Table 9: Prediction bias, standard deviation and root mean squared error under two compliance-weight models

λ	ϑ	Method	Weight model	Bias (MCSE)	SD	RMSE
0	0	I2SLS	Scaled logit	0.691 (0.005)	0.086	0.696
0	0	I2SLS	Spline logit	0.711 (0.005)	0.085	0.716
0	0	DML	Scaled logit	−0.025 (0.004)	0.077	0.081
0	0	DML	Spline logit	0.008 (0.005)	0.079	0.079
0	0.5	I2SLS	Scaled logit	0.675 (0.006)	0.100	0.682
0	0.5	I2SLS	Spline logit	0.716 (0.006)	0.098	0.723
0	0.5	DML	Scaled logit	−0.035 (0.005)	0.090	0.097
0	0.5	DML	Spline logit	0.028 (0.005)	0.093	0.097
0	1	I2SLS	Scaled logit	0.625 (0.008)	0.134	0.639
0	1	I2SLS	Spline logit	0.687 (0.007)	0.130	0.699
0	1	DML	Scaled logit	−0.065 (0.007)	0.120	0.137
0	1	DML	Spline logit	0.026 (0.007)	0.122	0.125
1	0	I2SLS	Scaled logit	0.543 (0.005)	0.084	0.549
1	0	I2SLS	Spline logit	0.564 (0.005)	0.083	0.570
1	0	DML	Scaled logit	0.129 (0.005)	0.085	0.155
1	0	DML	Spline logit	0.160 (0.005)	0.086	0.182
1	0.5	I2SLS	Scaled logit	0.511 (0.006)	0.098	0.520
1	0.5	I2SLS	Spline logit	0.557 (0.006)	0.096	0.565
1	0.5	DML	Scaled logit	0.106 (0.006)	0.101	0.146
1	0.5	DML	Spline logit	0.166 (0.006)	0.101	0.194
1	1	I2SLS	Scaled logit	0.446 (0.007)	0.126	0.464
1	1	I2SLS	Spline logit	0.521 (0.007)	0.123	0.535
1	1	DML	Scaled logit	0.064 (0.007)	0.128	0.143
1	1	DML	Spline logit	0.157 (0.007)	0.127	0.202

Notes. The two CATE settings fix $\kappa = 1$ and $\gamma_y = 0$, vary $\vartheta \in \{0, 0.5, 1\}$ and use 300 replications per row. Rows compare I2SLS and DML under the paper's scaled-logit weights and spline-logit weights. Naive 2SLS is absent because it uses no weights. Bias is mean prediction error, parentheses report SD divided by $\sqrt{300}$, SD is the standard deviation of prediction errors across replications, and RMSE is root mean squared prediction error. The table compares prediction performance, not weight-function accuracy, statistical significance or coverage. Results are conditional on the chosen calibration, and the design disclosures in Appendix I apply to this table.

difference $Y_i(1) - Y_i(0) = \tau_i$, so it does not affect $\tau^*(x)$. The step-function $g(X)$ can instead affect the estimated treatment-effect coefficients. Changes in those coefficients can change Terms I and II, and Term IV through the estimated slope. The direction and size of these changes depend on the other parts of the design. The corresponding column of Table 2 is zero up to numerical precision in every reported cell. The component is nonzero only when τ^* itself is nonlinear and the two complier distributions differ.

E.2 Performance on the Full Design Grid

These tables extend Section 4.2 to all 60 design combinations. The design varies the covariate scale, the instrument-propensity specification, $g(X)$ and selection into the target population. Each combination uses 500 replications. The step-function $g(X)$ leaves the conditional complier effect linear, with slope 10.

The estimator ranking depends on $g(X)$ and the covariate scale. With linear $g(X)$, DML has the lower RMSE of the two adjusted estimators in 26 combinations and I2SLS in four. With step-function $g(X)$, DML has the lower RMSE in 19 combinations and I2SLS in 11. The combinations in which I2SLS has the lower RMSE are at covariate scale 1, apart from three at scale 10 with $\vartheta \leq 0.25$. Naive 2SLS has the highest RMSE in 38 of the 60 combinations. In the other 22 the I2SLS RMSE is higher. The closest comparison under step-function $g(X)$ differs in RMSE by 0.004 (0.786 and 0.791 after rounding), so the ranking can represent a small performance difference.

Both adjusted estimators also have substantial negative bias at covariate scale 10 with the cubic instrument specification and no systematic selection. These results show that neither method performs well throughout the grid. Section 4.3 examines the components of prediction error, and Appendix D.4 examines the sensitivity to the covariate limits used for compliance weights.

Table 10: Monte Carlo prediction bias and root mean squared error, covariate scale 1

			Naive 2SLS			I2SLS			DML		
Instrument	Outcome	ϑ	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE
linear	linear	0	−0.645 (0.023)	−0.639	0.828	0.100 (0.021)	0.098	0.473	0.164 (0.022)	0.167	0.513
linear	linear	0.25	−3.061 (0.024)	−3.037	3.107	−0.175 (0.020)	−0.170	0.488	−0.091 (0.022)	−0.078	0.489
linear	linear	0.5	−5.183 (0.023)	−5.151	5.208	−0.389 (0.020)	−0.374	0.600	−0.280 (0.021)	−0.272	0.553
linear	linear	0.75	−7.034 (0.023)	−7.076	7.053	−0.588 (0.022)	−0.595	0.769	−0.439 (0.022)	−0.423	0.664
linear	linear	1	−8.590 (0.022)	−8.583	8.604	−0.788 (0.025)	−0.780	0.962	−0.601 (0.024)	−0.583	0.808
linear	step	0	−0.628 (0.025)	−0.646	0.839	0.109 (0.021)	0.097	0.485	0.151 (0.022)	0.146	0.512
linear	step	0.25	−3.028 (0.024)	−3.009	3.073	−0.041 (0.021)	−0.025	0.473	−0.072 (0.021)	−0.014	0.486
linear	step	0.5	−5.198 (0.024)	−5.217	5.225	−0.081 (0.022)	−0.130	0.506	−0.271 (0.022)	−0.300	0.553
linear	step	0.75	−7.073 (0.023)	−7.041	7.092	−0.112 (0.024)	−0.091	0.554	−0.483 (0.023)	−0.444	0.713
linear	step	1	−8.602 (0.022)	−8.609	8.615	−0.040 (0.026)	−0.029	0.585	−0.618 (0.025)	−0.618	0.829
cubic	linear	0	−0.767 (0.023)	−0.801	0.928	0.156 (0.021)	0.155	0.484	0.186 (0.022)	0.195	0.520
cubic	linear	0.25	−2.962 (0.023)	−2.957	3.005	−0.089 (0.021)	−0.095	0.472	−0.031 (0.021)	−0.004	0.471
cubic	linear	0.5	−5.001 (0.022)	−4.980	5.026	−0.365 (0.022)	−0.352	0.610	−0.235 (0.022)	−0.203	0.546
cubic	linear	0.75	−6.802 (0.021)	−6.789	6.819	−0.647 (0.024)	−0.638	0.839	−0.432 (0.023)	−0.452	0.676
cubic	linear	1	−8.264 (0.021)	−8.244	8.277	−0.868 (0.025)	−0.834	1.037	−0.590 (0.024)	−0.612	0.797
cubic	step	0	−0.984 (0.024)	−0.964	1.123	−0.016 (0.022)	0.009	0.488	0.215 (0.022)	0.229	0.544
cubic	step	0.25	−3.238 (0.023)	−3.196	3.281	−0.080 (0.022)	−0.085	0.496	−0.019 (0.021)	−0.031	0.472
cubic	step	0.5	−5.321 (0.023)	−5.311	5.346	0.003 (0.023)	−0.003	0.523	−0.259 (0.023)	−0.254	0.569
cubic	step	0.75	−7.102 (0.021)	−7.098	7.117	0.236 (0.025)	0.235	0.604	−0.450 (0.022)	−0.435	0.672
cubic	step	1	−8.567 (0.022)	−8.566	8.581	0.514 (0.027)	0.496	0.786	−0.574 (0.024)	−0.581	0.791

Notes. The table reports all 20 cells at covariate scale 1, with 500 replications per cell. Bias is mean prediction error relative to the true LATE of the realized target compliers, parentheses report its Monte Carlo standard error, median bias is the median prediction error across replications, and RMSE is root mean squared prediction error. Estimators, common compliance weights and the covariate limits used to calculate compliance weights are defined in Section 4.1.

Table 11: Monte Carlo prediction bias and root mean squared error, covariate scale 3

			Naive 2SLS			I2SLS			DML		
Instrument	Outcome	ϑ	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE
linear	linear	0	−4.525 (0.058)	−4.522	4.707	−0.171 (0.052)	−0.176	1.176	−0.015 (0.053)	−0.013	1.174
linear	linear	0.25	−20.770 (0.060)	−20.741	20.813	−1.453 (0.055)	−1.465	1.899	−0.880 (0.051)	−0.880	1.447
linear	linear	0.5	−30.511 (0.056)	−30.507	30.537	−2.562 (0.063)	−2.418	2.925	−1.595 (0.058)	−1.562	2.055
linear	linear	0.75	−35.787 (0.053)	−35.736	35.807	−3.052 (0.068)	−3.015	3.407	−1.886 (0.063)	−1.820	2.358
linear	linear	1	−38.817 (0.050)	−38.804	38.833	−3.064 (0.070)	−3.019	3.436	−1.819 (0.067)	−1.813	2.361
linear	step	0	−4.233 (0.066)	−4.225	4.480	0.265 (0.054)	0.222	1.238	0.020 (0.053)	−0.038	1.191
linear	step	0.25	−20.541 (0.059)	−20.535	20.583	1.877 (0.064)	1.858	2.357	−0.807 (0.053)	−0.837	1.437
linear	step	0.5	−30.409 (0.055)	−30.410	30.434	3.673 (0.064)	3.567	3.944	−1.663 (0.057)	−1.627	2.096
linear	step	0.75	−35.675 (0.052)	−35.694	35.694	4.771 (0.075)	4.701	5.058	−1.893 (0.065)	−1.800	2.390
linear	step	1	−38.694 (0.051)	−38.751	38.711	5.186 (0.078)	5.039	5.468	−1.895 (0.066)	−1.874	2.396
cubic	linear	0	−4.675 (0.061)	−4.683	4.869	1.128 (0.289)	1.199	6.546	1.345 (0.061)	1.341	1.916
cubic	linear	0.25	−17.474 (0.056)	−17.492	17.519	−3.953 (5.851)	1.694	130.763	0.160 (0.060)	0.264	1.344
cubic	linear	0.5	−25.206 (0.056)	−25.195	25.236	0.652 (6.459)	−1.966	144.288	−0.566 (0.062)	−0.622	1.496
cubic	linear	0.75	−29.601 (0.050)	−29.626	29.622	36.341 (27.096)	−7.988	606.378	−1.171 (0.066)	−1.181	1.888
cubic	linear	1	−32.224 (0.050)	−32.248	32.244	−9.113 (3.651)	−7.229	82.063	−1.110 (0.070)	−1.105	1.922
cubic	step	0	−3.944 (0.061)	−3.967	4.176	−0.862 (0.894)	0.574	19.978	1.406 (0.061)	1.379	1.958
cubic	step	0.25	−16.924 (0.057)	−16.807	16.971	10.462 (17.017)	−14.490	380.268	0.182 (0.058)	0.270	1.303
cubic	step	0.5	−24.836 (0.054)	−24.795	24.865	143.337 (126.957)	−31.231	2839.621	−0.727 (0.064)	−0.609	1.609
cubic	step	0.75	−29.293 (0.050)	−29.266	29.315	−277.890 (254.721)	50.222	5696.819	−1.209 (0.065)	−1.174	1.885
cubic	step	1	−31.963 (0.048)	−31.957	31.981	85.322 (15.210)	46.976	350.309	−1.296 (0.069)	−1.352	2.016

Notes. The table reports all 20 cells at covariate scale 3, with 500 replications per cell. Bias, its Monte Carlo standard error in parentheses, median bias, RMSE, the estimators, weights and covariate limits are defined as in Table 10.

Table 12: Monte Carlo prediction bias and root mean squared error, covariate scale 10

			Naive 2SLS			I2SLS			DML		
Instrument	Outcome	ϑ	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE	Bias (MCSE)	Median bias	RMSE
linear	linear	0	−36.379 (0.262)	−36.232	36.846	−14.792 (0.230)	−15.078	15.658	−14.142 (0.240)	−14.393	15.121
linear	linear	0.25	−102.252 (0.235)	−102.396	102.386	−0.423 (20.479)	2.405	457.461	−0.516 (0.220)	−0.334	4.938
linear	linear	0.5	−112.489 (0.202)	−112.400	112.579	−4.947 (4.750)	−7.702	106.216	1.330 (0.221)	1.288	5.114
linear	linear	0.75	−115.944 (0.190)	−115.848	116.021	−22.186 (13.358)	−6.970	299.219	1.035 (0.270)	0.745	6.114
linear	linear	1	−117.343 (0.181)	−117.499	117.412	−5.312 (3.786)	−6.063	84.736	1.089 (0.405)	0.393	9.104
linear	step	0	−32.482 (0.263)	−32.312	33.010	−10.766 (0.230)	−10.703	11.925	−13.726 (0.228)	−13.630	14.641
linear	step	0.25	−98.849 (0.226)	−98.972	98.978	−50.328 (54.812)	−62.708	1225.440	−1.333 (0.208)	−1.203	4.827
linear	step	0.5	−109.999 (0.209)	−109.954	110.098	118.730 (34.255)	83.157	774.357	0.619 (0.194)	0.502	4.379
linear	step	0.75	−113.902 (0.191)	−113.968	113.982	113.140 (46.100)	58.051	1035.999	0.349 (0.198)	0.556	4.437
linear	step	1	−115.837 (0.179)	−115.546	115.906	65.295 (33.124)	41.954	742.805	−0.269 (0.222)	−0.571	4.965
cubic	linear	0	−41.850 (0.374)	−41.815	42.676	−27.125 (0.225)	−27.065	27.588	−25.918 (0.239)	−25.987	26.461
cubic	linear	0.25	−107.735 (0.304)	−107.783	107.949	−1.244 (0.254)	−1.810	5.800	−3.400 (0.213)	−3.546	5.845
cubic	linear	0.5	−110.434 (0.286)	−110.147	110.618	−10.246 (10.236)	3.478	228.875	−0.577 (0.225)	−0.830	5.051
cubic	linear	0.75	−110.650 (0.249)	−109.998	110.790	6.519 (5.583)	2.555	124.895	0.400 (0.252)	0.632	5.642
cubic	linear	1	−111.026 (0.223)	−110.976	111.138	28.747 (33.443)	−1.895	747.619	1.489 (0.251)	1.482	5.801
cubic	step	0	−32.465 (0.370)	−32.855	33.501	−16.769 (0.236)	−16.866	17.576	−26.422 (0.233)	−26.434	26.931
cubic	step	0.25	−98.060 (0.328)	−97.847	98.334	−27.874 (1.976)	−22.753	52.201	−3.897 (0.204)	−3.951	5.993
cubic	step	0.5	−102.536 (0.289)	−102.640	102.739	−67.651 (118.225)	−94.464	2641.822	−0.893 (0.222)	−0.860	5.044
cubic	step	0.75	−104.782 (0.261)	−104.831	104.944	−256.315 (244.944)	−102.877	5477.640	−0.358 (0.248)	−0.408	5.553
cubic	step	1	−106.408 (0.230)	−106.430	106.533	95.143 (565.143)	124.417	12624.697	−0.198 (0.261)	−0.281	5.837

Notes. The table reports all 20 cells at covariate scale 10, with 500 replications per cell. Bias, its Monte Carlo standard error in parentheses, median bias, RMSE, the estimators, weights and covariate limits are defined as in Table 10.

F Complier Rate Diagnostics

This appendix explains how the frequency of compliers affects the information available for estimating their treatment effect. Write $P_C = \Pr(T_i(1) > T_i(0))$ for a population's complier share. In the Monte Carlo design, the draw C_i uses the scaled logistic probability in equation (31), with multiplier 0.4. A unit is a complier when $C_i = 1$. Always-takers are drawn only among units with $C_i = 0$, at rate $p_{AT}(x)/\{1-p_C(x)\}$, so the complier probability conditional on $X_i = x$ is $p_C(x)$. It averages over the latent draw ν_i. In the Card design, the reference population has a complier share of approximately 0.233632. Selection based on covariates can change this share in the sample and target populations.

Under the conditional IV assumptions in Assumption 1, treatment probabilities in the sample population satisfy

$$\mathbb{E}_S[T_i \mid Z_i = 1, X_i = x] - \mathbb{E}_S[T_i \mid Z_i = 0, X_i = x] = \pi_S(x).$$

Combining this identity with the binary-instrument covariance identity in Appendix B gives

$$\mathrm{Cov}_S(T_i, Z_i \mid X_i = x) = \pi_S(x)h_0(x)\{1 - h_0(x)\}.$$

This covariance is small when compliers are rare or when one instrument value is rare at the covariate values being considered.

For a conditional effect estimated as a ratio of covariances, a small denominator can magnify estimation error. Holding the denominator fixed, an error in the numerator is multiplied by its inverse. Estimating the denominator introduces additional uncertainty. The linear OrthoIV model instead combines information across covariate values through the weighted equations in Appendix B. Its precision depends on the distribution of compliance, instrument variation and covariates, as well as outcome variability. An overall complier share alone therefore does not determine its variance.

The Monte Carlo multiplier is fixed across combinations, but the covariate distribution changes. Consequently, fixing the multiplier does not hold average compliance or estimation precision fixed. The Card calibration increases the reference complier share to make the comparisons more informative at the chosen sample size, as explained in Section 5.1.

G Card Simulation Calibration Details

This appendix records the numerical specifications used in Section 5.1. Following the Census-division ordering used by Blandhol et al. (2026), $X_2 = 0, 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8, 1$ corresponds, respectively, to Mountain, South Atlantic, West South Central, West North Central, Pacific, Mid-Atlantic, New England, East North Central, and East South Central.

G.1 Instrument Probabilities

The coefficients in equation (36) are taken from the replication package of Blandhol et al. (2026), which estimates the polynomial using the Card data:

$$\left(c_{jk}\right)_{j,k=0}^{3} = \begin{pmatrix} 1.07 & -2.71 & 7.61 & -5.87 \\ -1.96 & 6.69 & -10.64 & 8.32 \\ 1.72 & -1.91 & -1.65 & -3.19 \\ 0.21 & -8.30 & 20.76 & -9.71 \end{pmatrix}. \tag{46}$$

The rows index powers of X_1, and the columns index powers of X_2, following the replication package. The probabilities are restricted to $[0.01, 0.99]$; no covariate combination in the calibration reaches either bound.

G.2 Compliance Types and Treatment Assignment

The initial compliance model implies an overall complier share of 0.077020. During calibration development, we considered several compliance models and selected a common overall

complier share that could be attained while limiting the covariate probability mass affected by the upper bound in equation (37) to at most 1% for each model. The initial target share of 0.30 was not attainable under this restriction. The resulting common share was 0.233632. For the compliance model retained in this paper, the corresponding scale is approximately 3.0334 (3.0334064206 before rounding), and the upper bound does not bind at any covariate value used in the calibration. Thus, scaling increases the level of compliance without changing the normalized complier covariate distribution. The common-share rule holds average compliance at the same level across the candidate models. Estimation precision can still differ because the distribution of compliance across covariates differs.

The probabilities $p_{T|Z=0}(x)$ and $p^{\zeta}_{T|Z=1}(x)$ are defined in Section 5.1. A uniform draw U_i independent of (X_i, Z_i) then assigns types, with always-takers at $U_i \leq p_{T|Z=0}(X_i)$, compliers between $p_{T|Z=0}(X_i)$ and $p^{\zeta}_{T|Z=1}(X_i)$, and never-takers above. Treatment is $T_i = \mathbf{1}\{U_i \leq p^{\zeta}_{T|Z=1}(X_i)\}$ when $Z_i = 1$ and $T_i = \mathbf{1}\{U_i \leq p_{T|Z=0}(X_i)\}$ when $Z_i = 0$, so monotonicity holds by construction.

G.3 Outcome Calibration

The outcome model in equation (38) contains 96 coefficients, with a separate 16-term polynomial for each compliance type and potential treatment status. We add the same error $\varepsilon_i \sim \mathcal{N}(0, 0.2)$ to both potential outcomes of individual i, where 0.2 is the variance. We select the coefficients using our implementation of the constrained optimization problem in the Blandhol et al. (2026) replication package. The objective is the sum of two squared differences. The first is between the mean log wage of the treated Card observations without a four-year college in the county and the mean of the treated always-taker outcome under the reference covariate distribution. The second is between the mean log wage of the untreated Card observations with a four-year college in the county and the mean of the untreated never-taker outcome under the same covariate distribution. These two model means average over the reference covariate distribution without additional weights for the probability of

each compliance type.

The equality constraints are the five calibration targets in Section 5.1, the restriction that the type-share-weighted sum of the nonlinear basis coefficients is zero for each t and each of the 13 nonlinear basis terms, and the restriction $\theta_{\text{never-taker},1} = \theta_{\text{never-taker},0}$.

Three families of inequalities hold at every support point. The level bound is $4.6 \leq \varphi(x)^\top \theta_{g,t} \leq 7.79$ for every (g,t), the effect bound is $|\varphi(x)^\top(\theta_{g,1} - \theta_{g,0})| \leq 2$ for every g, and the always-taker effect satisfies $\varphi(x)^\top(\theta_{\text{always-taker},1} - \theta_{\text{always-taker},0}) \geq 0$.

The accepted solution matches each of the five targets with an absolute equality residual of at most 10^{-8}.

H Comparison of Treatment-Effect Models in the Card Simulation

We examine whether allowing a nonlinear treatment-effect model improves DML prediction. We hold the estimated nuisance functions and compliance weights fixed and replace the baseline linear model with two cubic specifications. The total-degree cubic includes terms whose combined power in the two covariates is at most three. The tensor-product cubic allows powers up to three in each covariate, including interactions such as $X_1^3 X_2^3$. Both allow nonlinear treatment effects, with more terms in the tensor-product specification.

We compare these models using the instrument probabilities calibrated to the Card data ($\kappa = 1$), with $\gamma_y = 0$. The comparison includes both linear and nonlinear CATEs and all three selection settings.

Table 13 shows that neither cubic model improves RMSE over the linear model in any of the six settings. Their estimated absolute biases are smaller in some settings, but their prediction errors vary much more across replications.

For example, with a nonlinear CATE ($\lambda = 1$) and moderate selection ($\vartheta = 0.5$), the total-degree cubic has an estimated bias of 0.034, compared with 0.106 for the linear model.

Table 13: Bias, Standard Deviation, and RMSE by Treatment-Effect Model

ϑ	Treatment-effect model	Bias (MCSE)	SD	RMSE
Panel A: $\lambda = 0$				
0	Linear	−0.025 (0.004)	0.077	0.081
	Tensor-product cubic	−0.054 (0.134)	2.328	2.325
	Total-degree cubic	−0.002 (0.039)	0.667	0.666
0.5	Linear	−0.035 (0.005)	0.090	0.097
	Tensor-product cubic	−0.006 (0.227)	3.935	3.929
	Total-degree cubic	−0.062 (0.111)	1.927	1.925
1	Linear	−0.065 (0.007)	0.120	0.137
	Tensor-product cubic	−0.348 (0.244)	4.220	4.227
	Total-degree cubic	−0.108 (0.050)	0.861	0.867
Panel B: $\lambda = 1$				
0	Linear	0.129 (0.005)	0.085	0.155
	Tensor-product cubic	0.200 (0.184)	3.179	3.180
	Total-degree cubic	0.107 (0.031)	0.539	0.549
0.5	Linear	0.106 (0.006)	0.101	0.146
	Tensor-product cubic	0.281 (0.240)	4.150	4.152
	Total-degree cubic	0.034 (0.100)	1.732	1.730
1	Linear	0.064 (0.007)	0.128	0.143
	Tensor-product cubic	−0.188 (0.287)	4.966	4.961
	Total-degree cubic	0.079 (0.061)	1.048	1.049

Notes. 300 replications per row, $N = 18{,}000$ per draw, with $\kappa = 1$ and $\gamma_y = 0$. The three rows of each block share the same nuisance models and the same compliance weights, and they differ only in the treatment-effect model. Bias is the mean prediction error across replications, measured against that replication's own true target LATE, with its Monte Carlo standard error in parentheses. SD is the standard deviation of the prediction error across replications and RMSE is the root mean squared prediction error.
SD uses divisor $R - 1$ and RMSE uses divisor R, so RMSE can be below SD when bias is small. Results are conditional on the chosen calibration. Appendix I states its design limits.

However, the standard deviation of its prediction error is 1.732, compared with 0.101 for the linear model. Its RMSE is consequently much higher, at 1.730 compared with 0.146. The estimated bias of the cubic model is also less precisely measured, with a Monte Carlo standard error of 0.100.

The tensor-product cubic produces still greater variability and higher RMSE than the total-degree cubic in every setting. Thus, adding more nonlinear terms does not deliver more accurate target LATE predictions in this comparison.

These results show why flexibility in the treatment-effect model must be assessed together with estimation precision. A model that can represent a nonlinear CATE can still produce unstable predictions when its coefficients are estimated from the available sample. At the sample size considered here, the linear model has the lowest RMSE among the three specifications, including when the true CATE is nonlinear.

I Coverage of Nominal 95% Intervals in the Card Design

I.1 Design and Interval Construction

This appendix asks how often the reported 95% intervals contain the true target LATE in the Card simulation. The designs are defined in Section 5. The comparison between the two CATE settings varies CATE nonlinearity and selection. The additional comparisons vary the instrument probabilities or γ_y while holding the CATE linear.

Each interval is the prediction plus or minus 1.96 reported standard errors. Coverage is the fraction of replications in which this interval contains that replication's true target LATE. For I2SLS and DML, the standard errors condition on the estimated compliance weights and the observed target covariates. They omit uncertainty from estimating the weights, its covariance with coefficient estimation, and variation in the target covariate sample. The tables allow us to compare the reported standard errors with the actual variation in prediction errors across replications.

The calibration also leaves choices about treatment-effect heterogeneity. Neither the Card data nor the calibration targets uniquely determine the conditional complier effect or the aggregation gap G. Under the calibration constraints, the computed feasible ranges of G at $\vartheta = 0$ are $[-0.203187, 0.352505]$ at $\lambda = 0$ and $[-0.538863, 0.212275]$ at $\lambda = 1$. The adopted design gives $G = 0.121787$ and $G = -0.083433$, respectively. We use this same

calibrated outcome specification across the reported simulation blocks.

The bounds imposed when calibrating the nonlinear effect ($\lambda = 1$) need not continue to hold after replacing it with its linear approximation ($\lambda = 0$). The conditional mean treated outcome of compliers can then lie outside $[4.6, 7.79]$, and the absolute conditional effect can exceed 2. The share of compliers in the region violating these bounds is between 0.414% and 0.423% across selection settings. The linearized conditional effect ranges from -2.204 to 2.646. These bounds concern conditional means, and we do not clip the outcomes or effects after linearization.

I.2 Coverage Results

Tables 14 and 15 report coverage over 300 replications per setting. Coverage falls below the nominal 95% level for all reported DML settings. For naive 2SLS and I2SLS, several settings have no intervals containing the true target LATE.

Large bias explains an important part of these failures. Seventeen of the 54 reported rows have zero coverage, all for naive 2SLS or I2SLS with $\kappa = 1$ or $\kappa = 2$. In each of those rows, bias exceeds five times the root mean square reported standard error (rms SE). Intervals centered on such biased predictions can miss the target even when their estimated standard errors are otherwise informative about sampling variation.

Small bias does not ensure nominal coverage either. In the linear-CATE settings, DML's absolute bias is at most 0.072, but coverage ranges from 78.0% to 85.3%. Across all 18 DML rows, rms SE is about 70–83% of the empirical standard deviation of prediction errors. The reported standard errors therefore understate the observed variation. The omitted sources of uncertainty described above are possible contributors, but these tables do not identify their separate contributions.

With a nonlinear CATE, DML also has positive bias and coverage ranges from 50.7% to 84.0%. Assessing interval performance thus requires attention to both prediction bias and the uncertainty included in the standard error.

Table 14: Coverage of nominal 95% intervals as λ and ϑ vary

λ	ϑ	Method	Coverage (MCSE)	Bias	SD	rms SE
0	0	Naive 2SLS	0.000 (0.000)	0.744	0.088	0.086
0	0	I2SLS	0.000 (0.000)	0.691	0.086	0.070
0	0	DML	0.827 (0.022)	−0.025	0.077	0.058
0	0.5	Naive 2SLS	0.000 (0.000)	0.812	0.098	0.099
0	0.5	I2SLS	0.000 (0.000)	0.675	0.100	0.087
0	0.5	DML	0.820 (0.022)	−0.035	0.090	0.069
0	1	Naive 2SLS	0.000 (0.000)	0.853	0.118	0.117
0	1	I2SLS	0.000 (0.000)	0.625	0.134	0.112
0	1	DML	0.783 (0.024)	−0.065	0.120	0.088
1	0	Naive 2SLS	0.000 (0.000)	0.601	0.088	0.084
1	0	I2SLS	0.000 (0.000)	0.543	0.084	0.066
1	0	DML	0.507 (0.029)	0.129	0.085	0.066
1	0.5	Naive 2SLS	0.000 (0.000)	0.677	0.099	0.098
1	0.5	I2SLS	0.000 (0.000)	0.511	0.098	0.080
1	0.5	DML	0.687 (0.027)	0.106	0.101	0.079
1	1	Naive 2SLS	0.000 (0.000)	0.748	0.116	0.116
1	1	I2SLS	0.017 (0.007)	0.446	0.126	0.102
1	1	DML	0.840 (0.021)	0.064	0.128	0.101

Notes. The two CATE settings fix $\kappa = 1$ and $\gamma_y = 0$ and use 300 replications per row. Coverage is the share for which the estimate plus or minus 1.96 standard errors contains that replication's true target LATE. Parentheses report $\sqrt{p(1-p)/300}$, and 0.000 does not imply a known zero coverage probability. Bias is mean prediction error, SD is the standard deviation of prediction errors across replications, and rms SE is the root mean squared reported standard error. Naive 2SLS and I2SLS use HC1 standard errors, while DML uses HC0 from the OrthoIV treatment-effect regression. The adjusted-estimator standard errors condition on estimated weights and target covariates, omitting weight-estimation, coefficient-weight covariance and target-composition uncertainty. The reported intervals exhibit undercoverage. The table does not isolate the contributions of prediction bias and omitted sources of uncertainty to this shortfall. The design limits in this appendix apply.

Table 15: Coverage of nominal 95% intervals as κ and γ_y vary

κ	γ_y	ϑ	Method	Coverage (MCSE)	Bias	SD	rms SE
0	0	0	Naive 2SLS	0.870 (0.019)	0.073	0.089	0.086
0	0	0	I2SLS	0.823 (0.022)	0.001	0.086	0.061
0	0	0	DML	0.830 (0.022)	−0.012	0.076	0.055
0	0	0.5	Naive 2SLS	0.633 (0.028)	0.162	0.102	0.099
0	0	0.5	I2SLS	0.760 (0.025)	−0.052	0.101	0.071
0	0	0.5	DML	0.793 (0.023)	−0.024	0.090	0.064
0	0	1	Naive 2SLS	0.513 (0.029)	0.218	0.122	0.119
0	0	1	I2SLS	0.583 (0.028)	−0.140	0.118	0.089
0	0	1	DML	0.790 (0.024)	−0.066	0.109	0.080
2	0	0	Naive 2SLS	0.000 (0.000)	1.124	0.092	0.085
2	0	0	I2SLS	0.000 (0.000)	1.089	0.092	0.079
2	0	0	DML	0.853 (0.020)	−0.034	0.080	0.065
2	0	0.5	Naive 2SLS	0.000 (0.000)	1.176	0.098	0.098
2	0	0.5	I2SLS	0.000 (0.000)	1.089	0.108	0.100
2	0	0.5	DML	0.853 (0.020)	−0.051	0.095	0.079
2	0	1	Naive 2SLS	0.000 (0.000)	1.216	0.118	0.116
2	0	1	I2SLS	0.000 (0.000)	1.072	0.139	0.132
2	0	1	DML	0.780 (0.024)	−0.072	0.125	0.102
0	2	0	Naive 2SLS	0.893 (0.018)	0.062	0.095	0.094
0	2	0	I2SLS	0.880 (0.019)	0.007	0.094	0.075
0	2	0	DML	0.827 (0.022)	−0.013	0.077	0.055
0	2	0.5	Naive 2SLS	0.713 (0.026)	0.150	0.111	0.110
0	2	0.5	I2SLS	0.830 (0.022)	−0.058	0.110	0.087
0	2	0.5	DML	0.800 (0.023)	−0.024	0.090	0.064
0	2	1	Naive 2SLS	0.637 (0.028)	0.204	0.129	0.132
0	2	1	I2SLS	0.657 (0.027)	−0.161	0.126	0.108
0	2	1	DML	0.787 (0.024)	−0.066	0.109	0.081
0	10	0	Naive 2SLS	0.937 (0.014)	0.018	0.273	0.257
0	10	0	I2SLS	0.923 (0.015)	0.030	0.270	0.249
0	10	0	DML	0.833 (0.022)	−0.013	0.077	0.056
0	10	0.5	Naive 2SLS	0.940 (0.014)	0.104	0.291	0.299
0	10	0.5	I2SLS	0.947 (0.013)	−0.081	0.285	0.289
0	10	0.5	DML	0.793 (0.023)	−0.024	0.090	0.064
0	10	1	Naive 2SLS	0.937 (0.014)	0.145	0.335	0.361
0	10	1	I2SLS	0.907 (0.017)	−0.245	0.349	0.357
0	10	1	DML	0.793 (0.023)	−0.066	0.109	0.081

Notes. These blocks fix $\lambda = 0$ and use 300 replications per row. At $\kappa = 1$ the propensity matches Table 14, so the $\kappa = 0$, $\gamma_y = 0$ block is a different design from its $\lambda = 0$ block. Estimators, standard errors, coverage, the Monte Carlo standard error $\sqrt{p(1-p)/300}$ and interpretation limits are as in that table. A reported 0.000 does not imply a known zero coverage probability. The design limits in this appendix apply.