A Locally Robust Semiparametric Approach to Examiner IV Designs

Lonjezo Sithole *

April 21, 2026

Abstract

Examiner IV designs often require rich covariate adjustment. In such settings, the leniency instrument must be estimated, and current practice typically relies on unbiased jackknife IV (UJIVE), which is well suited to saturated linear first stages but restrictive when there are many examiners and many covariates. This paper develops a semiparametric estimator and asymptotic inference procedure for the standard covariate-adjusted examiner-IV estimand when the treatment propensity is estimated flexibly. I derive an examiner-specific orthogonal score, show that the associated Riesz representers admit explicit conditional-expectation formulas, and establish a multiple-robustness property of the resulting moment condition. Under mean-square consistency and product-rate conditions, the estimator is root-n consistent and asymptotically normal, and cross-fitting removes own-observation bias while reducing the effect of first-step estimation error on second-step inference. Monte Carlo evidence shows coverage close to nominal in a correctly specified benchmark design and sizable bias and RMSE improvements over naive plug-in estimation and linear UJIVE in a harder nonlinear design of the same sample size.

JEL Numbers: C14, C21, C26, C31, C45

Keywords: examiner IV, semiparametric estimation, orthogonal moment function, generated instruments, leniency designs, machine learning.

*Department of Economics, University of Michigan, 238 Lorch Hall, 611 Tappan Ave, Ann Arbor, MI 48109, USA

Email address: lsithole@umich.edu

1 Introduction

Examiner instrumental variable (IV) designs are increasingly popular in empirical economics. An early example of these examiner IV designs in applied economics is found in Kling (2006), who leverages random assignment of judges to estimate the effects of the duration of incarceration spells on two labor market outcomes: employment and earnings prospects. Since Kling (2006)'s seminal study, there has been a burst of literature exploiting idiosyncratic allocation of examiners to cases in pursuit of answers to a wide variety of interesting economic questions in settings beyond the criminal justice system.

The key idea in these designs is to exploit exogenous variation in the treatment propensity of these randomly assigned examiners in an IV strategy. For example, Kling (2006) uses the average incarceration length meted out by one's randomly assigned judge across all cases adjudicated by that judge as an IV for the potentially endogenous duration of incarceration spell. These designs differ from the textbook IV designs insofar as the examiner IV, the examiner's propensity to assign treatment status, is a latent variable that is estimated from data. To estimate the examiner IV, the canonical examiner IV design exploits jackknife methods to deal with "own observation bias", which arises when one's own treatment status is used in the estimation of the examiner's treatment propensity.

The standard approach to estimation and inference in examiner IV designs with a single binary endogenous treatment and covariates is the unbiased jackknife estimation method proposed by Kolesár (2013). In most settings, examiners have control over which times (days or time of day) they can work, and where or in which office they can work. Random assignment of examiners is, therefore, conditional on this sorting. To the extent that the random assignment is conditional, the treatment propensity is valid as an IV conditional on these covariates (conditionally ignorable). In some settings, the treatment propensity also varies across case characteristics while in others random assignment happens within batches of cases rather than at the individual level. Different applications, therefore, admit different configurations of covariates to cleanse the examiner IV of the residual confounding variation. Kolesár (2013)'s jackknife method allows for such covariates to avoid the well-known biases from canonical jackknife procedures in the presence of covariates.

However, Kolesár (2013)'s approach excludes a range of applications. His method hinges on the assumption that the approximation of the conditional expectation function for the treatment propensity is exactly linear in covariates. By construction, this linear approximation is exact in saturated specifications. In these specifications, empirical researchers necessarily include all the discrete covariates (e.g. strata indicators) and their interactions. With a small number of covariates, this approach works well but with even a moderate number of covariates, the dimension of the covariate vector scales up quite

quickly relative to the sample size. This happens even with routine controls. In the Philadelphia bail data analyzed by Coulibaly et al. (2024), there are eight judges and the reported controls include year, month, and day-of-week fixed effects. A saturated judge-specific first stage with judge effects and judge-by-control interactions then already implies 24 control dummies, 7 judge dummies, and $7 \times 24 = 168$ interaction terms, or roughly 199 coefficients before adding any defendant or offense controls. On the other hand, in non-saturated designs, including when continuous covariates are included to predict the treatment propensity of examiners, the assumption that the linear approximation is exact is restrictive and generally unwarranted. Instead, researchers may want to estimate the treatment propensity nonparametrically.

In general, existing methods are not suitable for applications where the number of covariates is quite large relative to the number of observations, including in the saturated specifications with high dimensional fixed effects. In such settings, it seems eminently reasonable to use machine learning (ML) methods to predict the treatment propensity. However, a growing recent literature suggests caution in use of ML in two-step estimation problems. This literature has established that plug-in estimators based on machine learning first steps introduce large regularization and model selection biases which contaminate estimation and inference in the second step (see, for example, Chernozhukov, Chetverikov, et al., 2018). In some settings, including applications involving naive LASSO plug-in estimators in two-step estimation, the two-step estimators are not even root-n consistent (Chernozhukov, Escanciano, et al. (2022)). I focus on these settings where empirical researchers estimate the conditionally ignorable treatment propensity nonparametrically with many examiners and possibly many covariates relative to the sample size.

I take as given the standard covariate-adjusted examiner-IV estimand that underlies the modern leniency-design literature: the Kolesár ratio formed with the generated leniency instrument $\gamma = \mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$. This paper then focuses on estimation of that examiner-IV target reliably when the treatment propensity cannot credibly be treated as saturated linear and instead must be learned flexibly with many examiners and many covariates. To answer this question, I develop a locally robust semiparametric estimator based on an examiner-specific orthogonal score. The score combines the identifying moment for the Kolesár estimand with an influence-function adjustment that removes the first-order effect of estimating the generated instrument. As I show below, this construction mitigates regularization and model-selection bias relative to naive plug-in approaches, addresses the own-observation bias that is central in examiner designs, and delivers root-n inference for the standard examiner-IV estimand under flexible first-stage estimation.

Contribution

The object of interest throughout is the covariate-adjusted examiner-IV estimand of Kolesár (2013). Under the standard assumptions used in leniency designs, this parameter recovers a convex average of treatment effects for units whose treatment status is shifted by quasi-random assignment to more or less lenient examiners. It is therefore the natural object of interest in the empirical literature when one wishes to retain the usual (weakly) causal interpretation of examiner variation while allowing for rich covariate adjustment; indeed, recent practitioner guidance recommends this leniency-design workflow and the corresponding UJIVE estimator as the preferred benchmark (Goldsmith-Pinkham et al., 2025). The contribution of this paper is a semiparametric estimation and inference framework for this empirically relevant examiner-IV estimand when the treatment propensity must be estimated flexibly in the presence of many examiners and many covariates.

This paper makes five concrete contributions to the literature on examiner IV designs and semiparametric estimation.

First, I derive the Neyman-orthogonal moment function for the Kolesár (2013) examiner IV estimand $\theta = \mathbb{E}[Y\gamma]/\mathbb{E}[T\gamma]$. While general recipes for constructing such moments exist (Chernozhukov, Escanciano, et al., 2022; Ichimura and Newey, 2022), the specific derivation for this examiner-IV ratio estimand with a generated instrument $\gamma = \mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$—involving two separate nuisance functions and their respective representers—is new. In this setting, the relevant representers can be written explicitly as conditional expectations, which yields an examiner-specific estimator rather than a purely generic application of the locally robust template. General frameworks such as Perez-Izquierdo (2026) treat ML-generated regressors in GMM but do not instantiate the examiner IV estimand or derive its specific representers.

Second, I establish that this orthogonal moment is multiply robust in the sense that, for each nuisance block $j \in \{1, 2\}$, valid estimation obtains when at least one of α_j or γ_j is correctly specified. With the explicit representers used here, correct specification of the corresponding outcome regressions is one sufficient route. This extends double robustness (as in Singh and Sun (2024) for binary instruments) to the many-instrument examiner IV setting.

Third, the framework permits a wide range of first-step estimators, including sieve methods and machine-learning procedures, provided they satisfy the mean-square consistency and product-rate conditions stated below. This is important in light of Blandhol et al. (2025), who show that once covariates enter the specification, conventional TSLS generally need not recover a convex average of causal effects. A flexible first stage is

closely tied to preserving the causal interpretation of the leniency-design estimand, rather than serving only as a statistical convenience.

Fourth, cross-fitting simultaneously eliminates own-observation bias (a key concern in examiner designs, cf. Angrist and Frandsen (2022)), reduces first-stage estimation error (Hansen and Kozbur, 2014; Jochmans, 2023), and obviates Donsker conditions that are not known to hold for ML estimators.

Fifth, relative to UJIVE (Kolesár, 2013; Goldsmith-Pinkham et al., 2025), this approach provides semiparametric estimation and inference for the same causal target without requiring the first stage to be exactly linear or saturated and with robustness to first-step misspecification. When a saturated linear first stage is a credible approximation, UJIVE remains a natural benchmark. The gain from the present approach is in settings where that approximation is restrictive because the examiner propensity is nonlinear or because the saturated specification becomes high-dimensional. Relative to Chao et al. (2023), who study linear IV regression with fixed or cluster-specific effects, many included exogenous regressors, and many weak instruments using jackknife-style estimators such as FELIM and FEFUL, my contribution is complementary: I study the Kolesár (2013) examiner-IV estimand as a semiparametric ratio estimand with a generated instrument and allow fully nonparametric or machine-learning first stages. I do not analyze the many-weak-instrument asymptotic problem considered by Chao et al. (2023). Relative to Wiemann (2023), the approach here handles high-dimensional covariates and provides formal multiple robustness guarantees.

Related Literature

This paper contributes to a growing theoretical literature on examiner IV designs. For comprehensive treatments of this literature, see the practitioner guides by Goldsmith-Pinkham et al. (2025) and Chyn et al. (2025). In important work, Frandsen, Lefgren, et al. (2023) revisit these examiner IV designs and propose a new approach to surmount violations of some of the assumptions underpinning standard approaches. Unlike Frandsen, Lefgren, et al. (2023), I do not consider the identification issues in this literature. Instead, I focus exclusively on estimation and inference in the presence of covariates. I build on the excellent work of Kolesár (2013) who proposes an unbiased jackknife estimation method that accounts for covariates in the estimation of the treatment propensity. While Kolesár (2013) contends that the treatment propensity can also be estimated nonparametrically using his method, his method does not account for the first-step bias that arises when nonparametric methods or machine learning are used in the estimation of the treatment propensity. Another closely related piece of work is

Mueller-Smith (2015) which, to the best of my knowledge, is the earliest application of machine learning to estimate the treatment propensity in examiner IV designs. Unlike Mueller-Smith (2015)'s approach, which uses LASSO under strong sparsity assumptions without adjusting for regularization and model selection biases, the approach I propose allows for a wide range of ML first steps, corrects for first-step estimation bias and delivers root-n consistent estimation that is not, in general, guaranteed for estimators based on ML first steps. The main examiner-specific implementation does not require strong sparsity assumptions and is robust to misspecification of some of the components estimated in the first step or the outcome model.

While I focus explicitly on examiner IV designs, this work also contributes to the broader IV literature. Particularly, it contributes to the optimal instruments literature with machine learning or nonlinear first steps, which spans the work of Belloni et al. (2012), Hansen and Kozbur (2014), Chernozhukov, Hansen, et al. (2015), Syrgkanis et al. (2019), Bai and Ng (2010), and Hartford et al. (2017). Belloni et al. (2012) and Chernozhukov, Hansen, et al. (2015) use LASSO under approximate sparsity assumptions, while Hansen and Kozbur (2014) and Bai and Ng (2010) use ridge regression and factor modelling respectively. Hartford et al. (2017) propose a deep learning approach to flexible counterfactual prediction in IV models. Syrgkanis et al. (2019) provides a more general framework for using ML to estimate conditional average treatment effects (CATE). Chen et al. (2021) provide justification for using ML in the first stage of linear IV models with sample-splitting, while Bruns-Smith (2025) develops a two-stage ML procedure for fully nonparametric IV regression. For the broader literature on instrumental variables with heterogeneous treatment effects, see the survey by Mogstad and Torgovitsky (2024). In this paper, I leverage locally robust semiparametric theory to provide a unified framework for semiparametric estimation of the IV model with treatment effect heterogeneity in Kolesár (2013) for a wide range of ML first steps in the presence of many exogenous covariates. This framework admits a completely nonparametric first step estimation of the "optimal" instrument in presence of many exogenous covariates under mild conditions.

More recent work closely related to this work is Wiemann (2023) who proposes using a version of K-means clustering to estimate the latent optimal instrument from a large set of candidate categorical instruments. While this approach can be applied to the examiner IV setting insofar as it accounts for many candidate categorical instruments, it does not deal with the aspect of many covariates that is relevant when the latent optimal instrument is conditionally ignorable and which motivates this work. On the linear-IV side, Chao et al. (2023) study IV regression with fixed or cluster-specific effects, many included exogenous regressors, and many weak instruments, and propose jackknife estimators such as FEJIV, FELIM, and FEFUL. Their contribution is complementary to mine: they solve a

many-weak-instruments problem in a linear IV framework, whereas I study the semiparametric examiner-IV ratio estimand with a generated instrument and orthogonal machine-learning first steps. Subsequent to the first arXiv version of this paper, Scheidegger et al. (2025) develop double machine learning inference for heterogeneous treatment effects using efficient machine-learning instruments. Their contribution is also complementary: they study a broader heterogeneous-effects IV setting with efficient instruments and kernel smoothing, whereas I focus on the Kolesár (2013) examiner-IV estimand with the generated leniency instrument $\gamma = \mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$ and many categorical examiners. Jochmans (2023) also proposes a group fixed effects characterization of the examiner fixed effects, inspired by the panel data literature. Like Wiemann (2023), Jochmans (2023) does not deal with issues arising from having many covariates. However, an important insight that I exploit from Hansen and Kozbur (2014) and Jochmans (2023) is that cross-fitting (or jackknife-type sample splitting) reduces the first stage estimation error in the construction of the latent instruments. Thus, besides dealing with "own observation bias" and lack of Donsker conditions which are not known to hold in high dimensional settings, the cross-fitting as applied in my framework also mitigates the first stage estimation error. Mikusheva (2021) also observes that cross-fitting reduces dependence between machine learning first steps and inference in the second step, which further bolsters the case for cross-fitting and alleviates concerns regarding possible distortions induced by machine learning in empirical practice that have been recently documented by Angrist and Frandsen (2022). Related practical work in the broader DML literature likewise emphasizes that moderate-sample performance can remain sensitive to sample splitting and first-step quality; see, for example, Zivich and Breskin (2021) and Ahrens et al. (2024).

This paper draws mainly from a recent but rapidly expanding literature on automatic debiased machine learning and locally robust semiparametric estimation (see Chernozhukov, Chetverikov, et al. (2018), Chernozhukov, Newey, Quintas-Martinez, et al. (2021), Chernozhukov, Newey, and Singh (2022a), Chernozhukov, Newey, and Singh (2022b), Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022)). Important antecedents on debiased inference in high-dimensional models in the statistics literature include Geer et al. (2014), Javanmard and Montanari (2014), and Zhang and Zhang (2014). Recent work has also identified important limitations of the standard DML approach: Hahn and Hausman (2021) show that in nonlinear control-variable models (including a version of the judge-leniency design), cross-fitting fails to eliminate many-instruments bias; Bonhomme et al. (2026) show that first-order Neyman orthogonality is insufficient for panel and network models with imprecisely estimated fixed effects, and propose higher-order orthogonalization; and Kolesár et al. (2025) demonstrate the fragility of sparsity-based estimators to reparametrization of the control matrix. As I discuss in Section 2, these

concerns operate differently in the present setting because the orthogonal moment function is affine in the first-step nuisance functions, so the relevant population-moment error admits an exact bilinear decomposition even though Neyman orthogonality itself remains a local derivative property. Chernozhukov, Chetverikov, et al. (2018) introduce various *double debiased* machine learning approaches to discipline use of machine learning in statistics and econometrics, with emphasis on reducing regularization and model selection biases from first step estimation. In this paper, I exploit the locally robust semiparametric approach that debiases the identifying moment function, which depends on a plug-in estimate of a first step nuisance function, by adding to it an influence function adjustment. In developing this approach, I primarily leverage theoretical insights from Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022). Using this approach, I provide an estimation method that allows for use of most of the off-the-shelf machine learning algorithms to estimate the examiner IV under mild regularity conditions in a setting where there are many examiners and possibly many covariates. The method also mitigates bias when other nonparametric methods are used, and allows for misspecification in the first step estimation, by virtue of a multiple robustness property of the orthogonal moment function.

The remainder of the paper is organized as follows: section 2 outlines and further motivates the estimation issues that arise in the examiner leniency design; section 3 presents the main theoretical results; section 4 reports Monte Carlo evidence; and section 5 concludes and provides directions for future work.

2 The Examiner Leniency Design: Estimation Issues in Empirical Practice

To fix concepts, I provide a quick overview of the examiner leniency design using an example from a canonical empirical setting: the US criminal justice system.

Suppose we are interested in the causal effect of pretrial detention on conviction (see Frandsen, Lefgren, et al. (2023)). Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment ($T_i = 1$, detained) and control ($T_i = 0$, released), respectively, so that the observed outcome is:

$$Y_i = Y_i(0) + (Y_i(1) - Y_i(0))T_i \tag{2.1}$$

The individual-level treatment effect, $Y_i(1) - Y_i(0)$, is allowed to vary across individuals. Here T_i is the bail decision (it is 1 when a defendant is detained and 0 otherwise); X_i is a vector of defendant and case characteristics; and Z_i is a $J \times 1$ vector of examiner (judge) indicators. The treatment status, T, is endogenous if there are unobservables that are correlated with

both treatment and the outcome. For example, well-to-do defendants can afford better lawyers and are therefore less likely to be detained and less likely to be convicted. Those who get treated (detained) are, therefore, potentially systematically different from those who do not get treated, so that a naive comparison of outcomes across treatment groups does not recover causal effects.

Remark 2.1 (Constant treatment effects as a special case)**.** *Under the restrictive assumption of constant treatment effects,* $Y_i(1) - Y_i(0) = \delta$ *for all* i*, equation (2.1) reduces to* $Y_i = \delta T_i + \mathbf{X}_i'\beta + \epsilon_i$ *for suitable* β *and* ϵ_i*. All results in this paper hold under this special case, but the framework accommodates the more general setting of heterogeneous treatment effects.*

The idea behind the examiner IV design is to instrument the endogenous treatment status, T_i, with one's assigned judge's propensity to detain or release, $\mathbb{E}[T_i \mid Z_i]$, where Z_i is a $J \times 1$ vector of examiner indicators. Since the examiners (judges, in this case) are quasi-randomly assigned to cases, this propensity to assign treatment status is independent of the unobservables. This treatment propensity, also known as the judge stringency or leniency measure (depending on how T is defined), is a valid instrument under the following assumptions: (i) the *exclusion restriction*, which requires that judges affect outcomes only through the bail decision; (ii) *monotonicity*; and (iii) *relevance*, which effectively requires that the examiners' treatment propensity varies non-trivially.

Monotonicity deserves careful discussion. In the examiner IV context, pairwise monotonicity requires that for any pair of judges j and j', if judge j is weakly more lenient than j', then every defendant released by j' would also be released by j. As noted by Imbens and Angrist (1994), this assumption can be violated when judges differ in their treatment criteria rather than simply in their overall stringency—for example, if one judge is lenient toward young defendants but strict toward older ones, while another judge exhibits the opposite pattern. Barua and Lang (2016) and Frandsen, Lefgren, et al. (2023) further examine such violations. Under the potential outcomes framework in equation (2.1), pairwise monotonicity ensures that the estimand from Kolesár (2013) yields a convex combination of pairwise local average treatment effects (LATEs). Frandsen, Lefgren, et al. (2023) also introduce the weaker assumption of *average first-stage monotonicity*—requiring monotonicity in treatment propensity on average across judges ordered by leniency, rather than defendant-by-defendant pairwise monotonicity—and Goldsmith-Pinkham et al. (2025) discuss this interpretation in the modern leniency-design framework; see also Coulibaly et al. (2024) for sharp testability of monotonicity in judge designs. I emphasize that my contribution is on estimation and inference for the standard examiner-IV ratio, conditional on the identification assumptions adopted by the researcher. The estimator itself does not depend on whether one invokes pairwise monotonicity or average first-stage monotonicity,

although the causal interpretation of the estimand does.

This principle of examiner (or judge) IV designs applies more generally in settings where there is an idiosyncratic assignment of individuals to a set of decision-makers, whose propensity to assign treatment status varies non-trivially across those decision-makers. It therefore comes as no surprise that examiner IV designs are being increasingly used for causal inference in economics and adjacent disciplines, given the wide range of settings where such idiosyncratic assignment of decision-makers takes place. These designs have been exploited to examine causal effects of foster care on various socioeconomic outcomes (e.g. Gross and Baron (2022) and Bald et al. (2022)); causal effects of bankruptcy on earnings, adverse financial events and foreclosure rates (e.g. Dobbie and Song, 2015); as well as causal effects of disability benefits on labor supply, household consumption and mortality (e.g. Black et al. (2018)), to pick but a few from the expanding list of applications.

Since Kling (2006)'s seminal study, several studies have used a leave-one-out sample analogue of $E[T_i \mid Z_i]$ as an examiner IV:

$$\hat{\gamma}_i^{\text{LOO}} = \frac{\sum_{i' \neq i} \mathbf{1}\{Z_{i'} = Z_i\} \tilde{T}_{i'}}{\sum_{i' \neq i} \mathbf{1}\{Z_{i'} = Z_i\}} \tag{2.2}$$

The leave-one-out approach is used to avoid "own observation bias" whereby one's own endogenous treatment status, if used in the estimation of the examiner IV, contaminates the instrument with the same endogeneity the design seeks to overcome. Moreover, in many applications, the assignment of examiners is not completely random. In the working example of the bail system in the US, judges have control over which courtrooms, days and shifts they work so there is clearly intentional sorting of judges into locations and times, which would otherwise invalidate the examiner IV. To cleanse the judge stringency measure of this confounding variation, the location-by-time fixed effects (and possibly other covariates) are partialled out from the treatment variable via an application of the Frisch-Waugh-Lovell theorem. Therefore, researchers use the residualized treatment, $\tilde{T}_i$, instead of T_i. With this estimate, they then apply the two stage least squares (TSLS). When the first stage is a saturated linear regression on examiner indicators after partialling out controls, this leave-one-out fitted value coincides with the jackknife IV (JIVE) first stage of Angrist, Imbens, et al. (1999).

An alternative approach that is currently recommended is Kolesár (2013)'s unbiased jackknife IV (UJIVE). In the examiner setting, the JIVE fitted value of Angrist, Imbens, et al. (1999) reduces to the leave-one-out examiner mean in a saturated first stage. UJIVE is different. It jackknifes the covariate-adjustment step itself and then uses the resulting

residualized first stage in the IV estimator. In the UJIVE approach of Kolesár (2013), the covariates are partialled out linearly via jackknife regressions under the assumption that the first stage regression is saturated so that a linear approximation of the conditional expectation function is exact.

The UJIVE formulation of Kolesár (2013) is based on the following decomposition of the conditionally ignorable treatment propensity with the covariate vector partialled out:

$$\gamma_i = \mathbb{E}(T_i \mid Z_i, X_i) - \mathbb{E}(T_i \mid X_i) \tag{2.3}$$
$$= Z_i'\beta_1 + X_i'\beta_2 - X_i'\alpha \tag{2.4}$$

where the second equality is the linear approximation of the conditionally ignorable treatment propensity.

What does "partialling out" mean in the nonlinear setting? The decomposition $\gamma = \mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$ is well-defined regardless of whether the conditional expectations are linear in their arguments. In the L^2 geometry, $\mathbb{E}[T \mid X]$ is the projection of T onto $L^2(X)$ (the space of square-integrable functions of X), while $\mathbb{E}[T \mid X, Z]$ is the projection of T onto $L^2(X, Z)$. Since $L^2(X) \subseteq L^2(X, Z)$, their difference γ lies in the orthocomplement of $L^2(X)$ within $L^2(X, Z)$, meaning $\mathbb{E}[\gamma \mid X] = 0$ and $\mathbb{E}[\gamma] = 0$ by the law of iterated expectations. The quantity γ thus captures the variation in the treatment propensity that is attributable to examiner assignment *net of* any variation explained by covariates alone. Under linearity, this reduces to the Frisch-Waugh-Lovell residual, but the L^2 characterization extends naturally to the fully nonparametric case. Note that this "partialling out" via the difference $\mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$ is conceptually distinct from using the partial derivative $\partial\mathbb{E}[T \mid Z, X]/\partial Z$, which would capture only marginal effects of Z and would not generally yield the same estimand.

Kolesár (2013)'s first-step estimator of γ_i, denoted $\hat{\gamma}_i$, is the sample analogue:

$$\hat{\gamma}_i = Z_i'\hat{\beta}_{1/i} + X_i'\hat{\beta}_{2/i} - X_i'\hat{\alpha} \tag{2.5}$$

where $\hat{\beta}_{1/i}$, $\hat{\beta}_{2/i}$ and $\hat{\alpha}$ are jackknife estimators (least squares estimators with observation i left out). A vector $\hat{\gamma}$ of the $\hat{\gamma}_i$'s is then plugged in as an instrument in a conventional IV regression in the second step. Operationally, this second step is just IV/2SLS with $\hat{\gamma}$ as the excluded instrument for T, using the same residualized outcome, treatment, and controls carried over from the first step. With covariates, the same residualized objects used in the first step are carried into the IV regression so that identification continues to come from

variation in examiner propensity net of the controls (see Section 6.2 of Kolesár (2013) for details).

First, note that since $\hat{\gamma}$ is an *estimate* of the true treatment propensity, the estimation error from the construction of γ, represented as $\hat{\gamma} - \gamma$, potentially contaminates inference in the second step (see Hahn and Ridder (2013)). Secondly, the construction above hinges crucially on the assumption that the linear approximation above is exactly linear in the examiner indicators, Z_i, and the covariate vector, X_i. With discrete controls, the first stage is saturated when all the controls and their interactions are included in the covariate vector, X. This often leads to high-dimensional fixed effects, which effectively "chop up" the data into cells and the examiner IVs are estimated in those cells. This potentially raises concerns of precision and many-weak IV (see Bhuller et al. (2020) for an example of this in practice).

In non-saturated specifications or when continuous covariates are included, the assumption of a perfectly linear conditional expectation function is dubious. As recently observed by Blandhol et al. (2025), when the first stage is not saturated and unless the first stage is estimated nonparametrically, the TSLS estimand does not yield a LATE within a convex hull of pairwise treatment effects. To use Blandhol et al. (2025)'s terminology, Kolesár (2013)'s estimand has a *weakly causal* interpretation when the first stage is perfectly linear or estimated nonparametrically. In non-saturated specifications or when continuous controls are included, Blandhol et al. (2025)'s insight suggests estimating the treatment propensity using some nonparametric methods such as sieves or polynomial splines. These methods, however, tend to produce biases which get transmitted to the second step where the parameter of interest is estimated. This motivates an approach that takes into account these biases in the second step.

More generally, specifications where the number of examiners and number of covariates are very large relative to sample size tip researchers over into the realm of high dimensional estimation and inference, where machine learning methods are routinely used. This paper provides a more general framework under mild assumptions which permit researchers to use an array of machine learning techniques to estimate treatment propensity in settings with large numbers of examiners and covariates that are not vanishingly small relative to the sample size. The framework also allows for first step nonparametric methods such as sieves or polynomial splines in settings with low to moderate numbers of examiners and controls.

It is worth noting that a growing body of work has documented challenges when combining machine learning methods with many fixed effects or many instruments. Angrist and Frandsen (2022) find that post-LASSO first stages exhibit substantial bias in overidentified IV models, and that split-sample IV and LIML outperform ML-based instrument selection. Kolesár et al. (2025) show that sparsity-based estimators are fragile to seemingly innocuous linear reparametrizations of the control matrix—such as which

examiner is the reference category when coding indicator variables—and that the sparsity assumption is often rejected empirically. In a nonlinear version of the judge-leniency design, Hahn and Hausman (2021) formally demonstrate that cross-fitting alone does *not* eliminate the many-instruments bias: the second-stage moment function's higher-order dependence on the nuisance parameters leaves residual bias terms that cross-fitting cannot remove. More generally, Bonhomme et al. (2026) show that first-order Neyman orthogonality can be insufficient when fixed effects are estimated imprecisely, as in panel data models with many units and short panels, and propose higher-order orthogonalization to address this. These findings raise a legitimate question: does the framework proposed here also suffer from these problems?

The structure of the proposed score addresses the particular bias mechanism emphasized by Hahn and Hausman (2021) and by work motivating higher-order orthogonalization. The orthogonal moment function

$$\psi = (Y - \theta T)\gamma + \alpha_1(T - \gamma_1) + \alpha_2(T - \gamma_2)$$

is *affine* in the first-step nuisance functions γ_1 and γ_2. Because of this affine dependence, the Gateaux derivative of ψ with respect to γ_j does not depend on γ_j itself, so the specific higher-order bias mechanism generated by nonlinear dependence of the score on the first steps, as in Hahn and Hausman (2021), does not arise here. Neyman orthogonality is still a local derivative property; what is exact in the present setting is the bilinear decomposition of the population-moment error. Theorem 3.1 relies on this exact structure: first-step estimation error enters the relevant population moment only through products of nuisance errors. Appendix B makes this explicit in Lemma B.1 and Corollary B.1. Theorem 4 of Chernozhukov, Escanciano, et al. (2022) then implies that, because the orthogonal moment is affine in the first steps, Neyman orthogonality and multiple robustness coincide. Under Assumption 3.2, the asymptotically relevant effect of first-step estimation error on second-step inference is therefore governed by this product term. This does not eliminate the need for good first-step estimation, a nondegenerate Jacobian, or caution about finite-sample performance; it clarifies the channel through which first-step error affects the asymptotics in the present setting.

The fragility-of-sparsity concern of Kolesár et al. (2025) is attenuated, though not eliminated, in the present setting. The orthogonal score requires only the product-rate condition in Assumption 3.2, so the asymptotically relevant effect of first-step error is second order: one of $\hat{\alpha}$ or $\hat{\gamma}$ may converge slowly provided the other converges sufficiently quickly. The framework is therefore not tied to LASSO. More generally, any first-step estimator satisfying Assumptions 3.1 and 3.2 is admissible, and methods that do not rely

on sparsity avoid the reference-category dependence emphasized by Kolesár et al. (2025).

This should not be read as a wholesale endorsement of machine-learning methods in high-dimensional fixed-effects problems. Poor first-step quality can still degrade finite-sample performance through slow convergence, weak first-stage variation, or noise in the estimated generated instrument. The concern in Angrist and Frandsen (2022) is also different from the one considered here: their negative findings arise from using LASSO to *select* which instruments enter the model, whereas the present approach uses flexible methods to estimate $\mathbb{E}[T \mid X, Z]$ with examiner indicators always included in the first step. Even so, finite-sample weakness of the estimated instrument remains a practical concern.

In light of these considerations, there is a natural practical recommendation for choosing first-step estimators. The fragility emphasized by Kolesár et al. (2025) is most acute for the treatment propensity functions $\gamma_1 = \mathbb{E}[T \mid X, Z]$ and $\gamma_2 = \mathbb{E}[T \mid X]$, where the examiner indicators in Z constitute many categorical regressors that interact with regularization in the problematic ways documented by those authors. For these functions, methods that do not rely on sparsity are often more attractive because they avoid dependence on a reference category choice. For the representers, the main estimator in this paper does not require a separate sparse regression, since it uses the explicit conditional-expectation identities derived below. If one instead adopts the automatic-Riesz route, LASSO-based Riesz regression remains available, but whether this is adequate should be verified in the application, particularly with respect to sensitivity to fixed-effect parameterization and tuning choices.

In the framework that I provide in the next section, I allow X and Z to be very large relative to the sample size, N. In such a high-dimensional setting, we can think of the dimensions of X and Z as *increasing with the sample size* (see Chernozhukov, Chetverikov, et al. (2018)). In the standard examiner IV designs, Z consists of examiner indicators. In my proposed approach, I allow for Z to be more general and the examiner fixed effects characterization arises as a special case. For example, Z can be examiner types, and therefore be identified with vectors of examiner's discrete and continuous characteristics such as age, race, experience and so on. Henceforth, in lieu of the linear approximation in equation (2.4) and motivated by the issues raised in this section, I use a nonparametric specification where the difference in conditional expectations is a difference in a square-integrable function of both X and Z and a square-integrable function of X only.

3 Orthogonal Estimation of the Kolesár Examiner-IV Estimand

In this section, I develop the theoretical results underlying the estimator. I first construct the orthogonal score by augmenting the identifying moment with a first-step influence-function adjustment. I then verify the orthogonality properties required by the locally robust semiparametric framework, including Neyman orthogonality and multiple robustness. Finally, I state high-level conditions under which the debiased estimator is asymptotically linear and asymptotically normal, and I provide a consistent variance estimator.

3.1 Construction of the Orthogonal Moment Function

The construction of the orthogonal moment function or locally robust score below follows Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022). The orthogonal score consists of two ingredients, namely: the identifying moment function (score) and the influence function adjustment (or correction term). The identifying moment function derives from the estimand of interest while the influence function adjustment follows from an orthogonality condition in a sense that I make precise later in this section.

I use the following IV estimand in Kolesár (2013):

$$\theta := \frac{\mathrm{Cov}(Y, \gamma)}{\mathrm{Cov}(T, \gamma)} = \frac{\mathbb{E}[Y\gamma]}{\mathbb{E}[T\gamma]} \tag{3.1}$$

where $\gamma := \mathbb{E}[T \mid X, Z] - \mathbb{E}[T \mid X]$ is the conditionally ignorable treatment propensity given a vector of covariates, X; T is a binary treatment status; Z is a vector of examiner indicators or examiner types; and Y is a scalar outcome variable. Informally, γ may be viewed as the treatment propensity after partialling out the covariates, X. More formally, treating each examiner as an instrument, γ can be interpreted as the strength of the instrument assigned to individual i relative to other potential instruments conditional on covariates (see Section 3.2 in Kolesár (2013)).

Kolesár (2013) shows that, under the potential outcomes framework (see equation (2.1)), this estimand yields a convex combination of local average treatment effects between pairs of judges under pairwise monotonicity, exclusion restriction and exogeneity. As discussed in Section 2, the weaker assumption of average first-stage monotonicity also suffices for a causal interpretation (Frandsen, Lefgren, et al., 2023; Goldsmith-Pinkham et al., 2025). This condition requires monotonicity in treatment propensity on average across judges

ordered by leniency, rather than defendant-by-defendant pairwise monotonicity. For that reason, the parameter serves as the covariate-adjusted causal estimand delivered by the standard leniency-design framework when examiner assignment shifts treatment propensity across otherwise comparable cases. This makes it an appropriate target in applications where rich covariate adjustment is indispensable. The relevance of this target is further reinforced by Blandhol et al. (2025), who show that once covariates are included, conventional TSLS generally need not admit a LATE interpretation. In this paper, I therefore take the Kolesár (2013) estimand as the causal parameter of interest and study how to estimate it reliably when the first step must be learned flexibly. Frandsen, Lefgren, et al. (2023) provide alternative estimators when the stronger versions of monotonicity and exclusion restriction do not hold, while Frandsen, Leslie, et al. (2023) provide another estimator that takes into account the fact that in some settings such as bail hearings in the US, batches of cases rather than individual cases are randomly assigned. My framework can accommodate the latter estimator under suitable conditions by simply including batch indicators in the covariate vector, X. It does not immediately extend to the estimator in Frandsen, Lefgren, et al. (2023), and I leave that extension to future work.

An oracle estimator based on the parameter above would simply be:

$$\hat{\theta} = \frac{\mathbb{E}_n[Y\gamma]}{\mathbb{E}_n[T\gamma]} \tag{3.2}$$

where $\mathbb{E}_n$ denotes the sample analogue of expectation and γ is treated as known. This estimator is infeasible because γ is unknown. A feasible estimator replaces the conditionally ignorable treatment propensity with an estimate and then plugs this estimate into the second step. Kolesár (2013) estimates the treatment propensity using an unbiased jackknife procedure, described in Section 2, to avoid own-observation bias. As discussed below, cross-fitting likewise eliminates own-observation bias while also dispensing with Donsker conditions. However, Kolesár (2013)'s approach excludes a growing class of applications in which the number of covariates and examiner fixed effects is large relative to the sample size, precisely the setting in which flexible nonparametric or machine-learning first steps are most attractive.

Define the identifying moment function $g(W, \gamma, \theta)$ by the condition $\mathbb{E}[g(W, \gamma, \theta)] = 0$ at $\theta = \theta_0$, where $W = (Y, T, X, Z)$. Rearranging Kolesár (2013)'s IV estimand yields

$$g(W, \gamma, \theta) = [Y - \theta T]\gamma. \tag{3.3}$$

To see this, notice that $\mathbb{E}[(Y - \theta T)\gamma] = \mathbb{E}[Y\gamma] - \theta\mathbb{E}[T\gamma] = 0$ whenever $\theta = \theta_0$. The validity of this moment condition also follows from instrument exogeneity: if γ is a valid

instrument, then the idiosyncratic error term in the outcome model should be orthogonal to the instrument, implying $\mathbb{E}[(Y - \theta T)\gamma] = 0$.

The next important ingredient for the orthogonal moment condition is the influence function adjustment. Use of influence functions to correct for biases in first step estimation has a long history in econometrics (see Newey (1994), Hahn and Ridder (2013), Hahn (1998), and Ai and Chen (2003)). The construction of the influence function adjustment in this section is based on the work of Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022), who provide a general form of the influence function, provided a certain "exogenous" orthogonality condition is satisfied. Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022) show that if the following "exogenous" orthogonality condition is satisfied:

$$E_F[\delta(X)\lambda(W, \gamma(F)(X))] = 0 \quad \text{for all} \quad \delta \in \Gamma \tag{3.4}$$

then under some regularity conditions, the influence function adjustment takes the following general form:

$$\phi(w, \gamma, \alpha, \theta) = \alpha(x, \theta)\lambda(w, \gamma(x)) \tag{3.5}$$

where $\lambda(w, \gamma(x))$ is a nonparametric residual function and $\alpha(x, \theta)$ is a special function called a Riesz representer.

In the present setting, the generated instrument is the difference of two conditional expectations. Let $\gamma_1 = \mathbb{E}[T \mid X, Z]$ and $\gamma_2 = \mathbb{E}[T \mid X]$, so that $\gamma = \gamma_1 - \gamma_2$. By definition of conditional expectation, γ_1 is the L^2 projection of the endogenous treatment variable T onto $\Gamma_1 = L^2(X, Z)$, while γ_2 is the L^2 projection of T onto $\Gamma_2 = L^2(X)$.

These projection characterizations deliver the orthogonality condition in equation (3.4). Specifically, because γ_1 is the projection of T onto $L^2(X, Z)$,

$$\mathbb{E}\big[\delta(X, Z)(T - \gamma_1(F)(X, Z))\big] = 0 \quad \text{for all } \delta \in L^2(X, Z),$$

and, analogously, because γ_2 is the projection of T onto $L^2(X)$,

$$\mathbb{E}\big[\delta(X)(T - \gamma_2(F)(X))\big] = 0 \quad \text{for all } \delta \in L^2(X).$$

Here I index γ_j by the underlying distribution F only to emphasize that these identities hold under the true data-generating process; I suppress this dependence below to avoid notational clutter. Consequently, γ_1 and γ_2 satisfy the "exogenous" orthogonality condition of Ichimura and Newey (2022) (equation 2.10 in Chernozhukov, Escanciano, et al. (2022)). In the notation

of Chernozhukov, Escanciano, et al. (2022), this condition can be written as

$$\mathbb{E}[\delta(X,Z)\lambda(W,\gamma_1(X,Z))] = 0 \quad \text{for all } \delta \in \Gamma_1, \text{where } \lambda(W,\gamma_1(X,Z)) = T - \gamma_1(X,Z).$$

Analogously:

$$\mathbb{E}[\delta(X)\lambda(W,\gamma_2(X))] = 0 \quad \text{for all } \delta \in \Gamma_2, \text{where } \lambda(W,\gamma_2(X)) = T - \gamma_2(X).$$

Accordingly, the generalized-regression theory in Chernozhukov, Escanciano, et al. (2022) applies here with modifications dictated by the identifying moment. Because γ comprises two first-step functions, γ_1 and γ_2, the influence-function adjustment likewise contains two correction terms, one for each nuisance regression (see Ichimura and Newey (2022), Chernozhukov, Escanciano, et al. (2022), and Newey (1994)). In the present setting, equation (3.5) therefore specializes to

$$\phi(W,\gamma,\alpha,\theta) = \alpha_1(X,Z)(T - \gamma_1(X,Z)) + \alpha_2(X)(T - \gamma_2(X)) \tag{3.6}$$

where $\alpha = (\alpha_1, \alpha_2)$ collects the two Riesz representers, α_1 and α_2; $\gamma = h(\gamma_1, \gamma_2)$ is a function of γ_1 and γ_2; and $\lambda(W,\gamma_1(X,Z)) = T - \gamma_1(X,Z)$ and $\lambda(W,\gamma_2(X)) = T - \gamma_2(X)$ are nonparametric residuals after projecting the endogenous treatment indicator, T, onto the respective spaces of square-integrable functions of (X,Z) and of X. Since T is binary, the true conditional expectations $\gamma_1(X,Z)$ and $\gamma_2(X)$ lie in $[0,1]$, although finite-sample approximations need not automatically respect these bounds. In practice, one may either use first-step estimators whose fitted values are constrained to $[0,1]$ or adopt a link-based specification such as a logistic regression. For the theory in this paper, I work with the representation in (3.6).

Remark 3.1 (Why two separate nuisance functions?)**.** *Since $\gamma_2(X) = \mathbb{E}[\gamma_1(X,Z) \mid X] = \int \gamma_1(X,z) f(z \mid X) dz$, one could in principle treat γ_1 as the single primitive nuisance function and derive γ_2 from it. This alternative would reduce the number of independently estimated nuisance components and could potentially weaken Assumption 3.1 by imposing consistency only on $\hat{\gamma}_1$. However, treating γ_1 and γ_2 as separate nuisance functions has important practical advantages. First, it preserves the multiple robustness property: the estimating equation remains valid if the outcome model or at least one of γ_1, γ_2 is misspecified. If γ_2 were derived from γ_1, misspecification of γ_1 would automatically contaminate γ_2, eliminating this robustness. Second, separate estimation allows researchers to use different first-step estimators (or tuning parameters) for each regression, exploiting the potentially different structures of $\mathbb{E}[T \mid X,Z]$ and $\mathbb{E}[T \mid X]$. Third, estimating γ_2 directly via a regression of T on X is computationally straightforward,*

whereas deriving γ_2 from γ_1 would require numerical integration over $Z \mid X$, which may be impractical with many examiner indicators.

Although closed-form expressions for the Riesz representers are available in the present conditional-expectation setting, it is still useful to discuss an estimation strategy that relies only on the orthogonal score and extends readily to other environments. In this paper, however, the main examiner-specific estimator uses the explicit conditional-expectation identities for the representers. Section 3.3.2 therefore discusses the "automatic" approach primarily as a more general construction.

Now that we have the identifying moment function and the influence function adjustment, we can write down the orthogonal score:

$$\psi(W, \gamma, \theta, \alpha) = [Y - \theta T]\gamma + \alpha_1(X, Z)(T - \gamma_1(X, Z)) + \alpha_2(X)(T - \gamma_2(X)) \tag{3.7}$$

Notice that this orthogonal score is a valid moment condition since it is zero in expectation. To see why, note that from our orthogonality argument above, $\mathbb{E}(\delta(X, Z)(T - \gamma_1(F)(X, Z))) = 0$ for all $\delta \in L^2(X, Z)$, and α_1 is an element in $L^2(X, Z)$ (more on this in the next section). It immediately follows that $\mathbb{E}(\alpha_1(X, Z)(T - \gamma_1(X, Z))) = 0$. Analogously, $\mathbb{E}(\alpha_2(X)(T - \gamma_2(X))) = 0$ by the fact that α_2 is an element in $L^2(X)$. The identifying moment function is zero in expectation, as shown earlier. Furthermore, as shown in Ichimura and Newey (2022) and Chernozhukov, Escanciano, et al. (2022), the orthogonal moment function (or score) is also an influence function. To distinguish between this influence function characterization of the orthogonal moment function and the influence function adjustment above, Ichimura and Newey (2022) and Chernozhukov, Escanciano, et al. (2022) refer to the latter as the "first step influence function" to emphasize its role as a bias correction term for the first step estimation.

3.2 Neyman Orthogonality and Multiple Robustness

The orthogonal moment function we have constructed ought to satisfy two orthogonality properties. The first property, known as Neyman orthogonality, requires that varying the first step function away from its true underlying value should have no effect locally on the average orthogonal moment function. More formally (see equation 2.4 in Chernozhukov, Escanciano, et al. (2022)):

$$\frac{d}{dt}\mathbb{E}\left[\psi\left(W, \gamma_0 + t\delta, \alpha_0, \theta\right)\right] = 0 \quad \text{for all } \delta \in \Gamma \text{ and } \theta \in \Theta \tag{3.8}$$

where γ_0 is the value of γ under the true underlying distribution of the data; δ represents a deviation of γ away from γ_0; the scalar, t, is the size of the deviation; and the derivative

above is evaluated at $t = 0$. In a setting with multiple steps like ours, this derivative is taken with respect to t for each first step function, holding the other first step functions fixed.

Let us hold γ_2 fixed. Consider the linear functional on $\Gamma_1 = L^2(X, Z)$ given by

$$\mathcal{L}_1(\delta) = \mathbb{E}[(Y - \theta T)\delta(X, Z)].$$

By the Riesz representation theorem, there exists a unique $\alpha_{01} \in \Gamma_1$ such that

$$\mathcal{L}_1(\delta) = \mathbb{E}[\alpha_{01}(X, Z)\delta(X, Z)] \quad \text{for all } \delta \in \Gamma_1.$$

In this conditional-expectation setting, the representer is $\alpha_{01}(X, Z) = \mathbb{E}[Y - \theta T \mid X, Z]$. Let $\gamma_{01}(X, Z) = \mathbb{E}[T \mid X, Z]$ denote the true first-step regression. It follows that:

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\psi\left(W, \gamma_{01} + t\delta, \alpha_0, \theta\right)\right] \\
&= \frac{d}{dt}\mathbb{E}\left[(Y - \theta T)\big((\gamma_{01} + t\delta) - \gamma_2\big) + \alpha_{01}\left(T - \gamma_{01} - t\delta\right) + \alpha_2\left(T - \gamma_2\right)\right] \\
&= \mathbb{E}\left[(Y - \theta T)\delta - \alpha_{01}\delta\right] = \mathbb{E}[((Y - \theta T) - \alpha_{01})\delta] = 0, \text{ since } \alpha_{01} = \mathbb{E}[Y - \theta T \mid X, Z].
\end{aligned} \tag{3.9}$$

An analogous calculation holding γ_1 fixed yields the second representer $\alpha_{02}(X) = -\mathbb{E}[Y - \theta T \mid X]$ and establishes Neyman orthogonality with respect to γ_2 as well.

Another property which must be satisfied is the following:

$$\mathbb{E}\left[\phi\left(W, \gamma_0, \alpha, \theta\right)\right] = 0 \text{ for all } \theta \in \Theta \text{ and } \alpha \in \mathcal{A}. \tag{3.10}$$

This follows from the projection argument at the end of the previous subsection.

Theorem 4 in Chernozhukov, Escanciano, et al. (2022) allows us to verify whether our orthogonal moment function is "multiply" robust. Multiple robustness, a generalization of double robustness from the case of one first step function to multiple first step functions, is a stronger requirement than Neyman orthogonality. Neyman orthogonality is a "local" property: varying the first step functions should not have an effect locally on the average moment function. Multiple (or double) robustness imposes a more stringent requirement: the outcome model or one or more of the first step functions may be misspecified, and the orthogonal score will still remain valid as an estimating equation. Theorem 4 of Chernozhukov, Escanciano, et al. (2022) provides a condition under which Neyman orthogonality and multiple (or double) robustness coincide, namely: the orthogonal moment function should be affine in the first step functions. Typically, verifying this condition entails inspecting the identifying moment function and the first step influence function adjustment. If both are affine in the first step function, then the orthogonal

moment function is affine in the first step function and the equivalence between Neyman orthogonality and multiple robustness follows immediately by Theorem 4.

We have already shown Neyman orthogonality and we can easily verify by inspection that both the identifying moment function and the influence function adjustment in our setting are affine in one first step function (say, γ_1), holding the other (say, γ_2) fixed. The multiple robustness of our orthogonal moment function, therefore, follows immediately by invoking Theorem 4 of Chernozhukov, Escanciano, et al. (2022). The implication of this result is that the orthogonal moment function proposed here is robust both to biases from first-step estimation and to misspecification of either some of the first-step components or the outcome model.

3.3 Estimation

3.3.1 The Method of Moments Estimator via Cross-Fitting

The method of moments estimator we present in this section combines the multiply robust orthogonal score constructed above with cross-fitting. The estimator is a root of the following debiased sample moments, which are the empirical analogue of population moment conditions based on the orthogonal score:

$$\begin{aligned}\hat{\psi}(\theta) = &\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}(Y_i - \theta T_i)\left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i)\right) \\ &+ \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\hat{\alpha}_{1l}(X_i, Z_i)\left(T_i - \hat{\gamma}_{1l}(X_i, Z_i)\right) \\ &+ \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\hat{\alpha}_{2l}(X_i)\left(T_i - \hat{\gamma}_{2l}(X_i)\right) = 0. \end{aligned} \tag{3.11}$$

For notational simplicity, I suppress the dependence of $\hat{\alpha}_{1l}$ and $\hat{\alpha}_{2l}$ on the fold-specific pilot $\tilde{\theta}_l$ in (3.11); throughout the asymptotic theory these representers are evaluated at $\tilde{\theta}_l$.

The double summation operator denotes a cross-fitting scheme, which generalizes sample splitting. The sample is partitioned into L folds, I_l for $l = 1, \ldots, L$. For each fold I_l, the observations in the complement ${I_l}^c$ are used to estimate the Riesz representers, α_{1l} and α_{2l}, and the first-step functions, γ_{1l} and γ_{2l}. The sample moment $\hat{\psi}$ then evaluates the orthogonal score on observations in the held-out fold using nuisance estimates learned on ${I_l}^c$. The debiased examiner IV estimator, $\hat{\theta}$, is any solution to these debiased sample moments.

Cross-fitting is valuable for several reasons. Most importantly, it obviates the need for Donsker conditions, which play a central role in classical semiparametric theory but are not

known to hold for high-dimensional machine-learning first steps (Chernozhukov, Chetverikov, et al., 2018; Chernozhukov, Newey, and Singh, 2022b; Chernozhukov, Escanciano, et al., 2022). As in sample splitting more generally (see Angrist, Imbens, et al. (1999)), cross-fitting also eliminates own-observation bias (Chernozhukov, Escanciano, et al., 2022), a prominent concern in examiner IV designs. In the present setting it further reduces bias arising from estimation of examiner treatment propensities (Jochmans, 2023). Finally, Mikusheva (2021) note that cross-fitting weakens the dependence between flexible first-step estimation of the instrument and second-step IV inference. That concern is less acute here because the generated instrument is learned from exogenous covariates, but the same logic reinforces the case for cross-fitting.

For the estimation of $\gamma_1 = \mathbb{E}[T \mid X, Z]$ and $\gamma_2 = \mathbb{E}[T \mid X]$, a range of regression methods may be used, provided the resulting first-step estimates satisfy the conditions imposed in Assumptions 3.1 and 3.2. As discussed in Section 2, methods that do not rely on sparsity are often attractive for γ_1 and γ_2 in the examiner-IV setting because they avoid the reference-category fragility documented by Kolesár et al. (2025). Specifically, γ_1 is estimated by regressing T on a dictionary of functions of (X, Z) (e.g., examiner indicators interacted with covariates), while γ_2 is estimated analogously by regressing T on a dictionary of functions of X only (dropping Z from the dictionary). The generated instrument is then $\hat{\gamma} = \hat{\gamma}_1 - \hat{\gamma}_2$. High-level rate results of this sort are available for many flexible first-step estimators; see Section 3.2.3 in Chernozhukov, Newey, Quintas-Martinez, et al. (2024) for references.

Turning to the Riesz representers, recall that $\alpha_1(X, Z) = \mathbb{E}[Y - \theta T \mid X, Z]$ as derived in Section 3.2. By the same Riesz representation argument applied to $\Gamma_2 = L^2(X)$, the second Riesz representer takes the form:

$$\alpha_2(X) = -\mathbb{E}[Y - \theta T \mid X] \tag{3.12}$$

The negative sign arises because γ_2 enters the identifying moment function with a negative sign (since $\gamma = \gamma_1 - \gamma_2$). To see this explicitly, holding γ_1 fixed, the Gateaux derivative of the identifying moment function with respect to γ_2 in direction δ yields $-\mathbb{E}[(Y - \theta T)\delta]$, so that the Riesz representer satisfying the orthogonality condition is $\alpha_2 = -\mathbb{E}[Y - \theta T \mid X]$.

The estimation of both Riesz representers is discussed in the next subsection. In the present setting they can be handled in two ways: by exploiting the explicit identities above, which is the main path to estimation in this paper, or by estimating them automatically from the orthogonal score. The next subsection discusses the latter generic approach.

3.3.2 Automatic Estimation of the Riesz Representers

This subsection discusses a generic automatic approach to estimation of the Riesz representers. In the present examiner-IV setting, the explicit examiner-specific identities are available and they provide the preferred implementation. The point of the discussion below is therefore not that the automatic approach is needed for the baseline estimator in this paper, but rather that it avoids repeated analytical derivations and extends naturally beyond the present conditional-expectation setting.

In principle, one may estimate a Riesz representer by plugging estimators into its closed-form expression whenever such an expression is available. In the present examiner-IV setting, these closed-form expressions are available and they underpin the main estimator studied in this paper. For the standard Kolesár (2013) estimand considered here, this closed-form route is therefore the natural one. In other problems, however, the closed-form argument can be cumbersome or numerically unstable: for conditional expectations the representer takes the form of a projection, while in settings such as average treatment effects, policy effects, or other extensions of the examiner-IV framework it may involve inverse propensity weights, density ratios, or other objects that are less transparent. For that reason, it is still useful to discuss the now-standard "automatic" approach, which constructs the representer directly from the orthogonal score and the data.

In moderate dimensional settings where the number of parameters increases slowly with the sample size, the Riesz representers can be approximated by "parametric sieves" and researchers may use estimation methods from the well-established literature on sieve estimation (see, for example, Chen (2007) for a handbook-chapter treatment of this broad literature on sieve estimation). For expositional purposes, I illustrate one such generic implementation with regularized least squares in a high-dimensional setting. In particular, I use LASSO as a convenient example for estimating the Riesz representers, focusing first on $\alpha_1(X, Z)$ (analogous arguments apply to $\alpha_2(X)$).

The population orthogonality condition for α_1 is

$$\mathbb{E}\left[((Y - \theta T) - \alpha_1(X, Z))\, \delta(X, Z)\right] = 0 \quad \text{for all } \delta \in L^2(X, Z),$$

which is satisfied when $\alpha_1(X, Z) = \mathbb{E}[Y - \theta T \mid X, Z]$. In the high-dimensional setting, I represent $\alpha_1(X, Z)$ as a linear combination of a dictionary of functions, $\rho_1' b(X, Z)$, where $b(X, Z) = (b_1(X, Z), \ldots, b_p(X, Z))'$ spans a rich approximation space. This leads to the regularized least-squares estimator

$$\hat{\alpha}_{1l}(x, z) = \hat{\rho}_1' b(x, z),$$

where, for a suitably chosen penalty level r_1,

$$\hat{\rho}_1 = \underset{\rho_1}{\operatorname{argmin}} \frac{1}{n - n_l} \sum_{i \notin I_l} \left((Y_i - \tilde{\theta}_l T_i) - \rho_1' b(X_i, Z_i) \right)^2 + 2 r_1 \sum_{j=1}^{p} |\rho_{1j}|. \tag{3.13}$$

Here $\tilde{\theta}_l$ is an initial fold-specific estimate of θ constructed on the auxiliary sample used to learn the nuisance functions.

To obtain this initial estimate, $\tilde{\theta}_l$, one may reserve an additional auxiliary split within the training sample for fold l and estimate the original identifying moment on observations that are not used for score evaluation; see Section 2.2 of Chernozhukov, Escanciano, et al. (2022) for a generic construction of this kind. I continue to write this pilot as $\tilde{\theta}_l$ to keep the notation consistent with the rest of the paper. As argued by Chernozhukov, Escanciano, et al. (2022), this initial estimate does not affect the asymptotic distribution of the debiased examiner IV estimator because of Neyman orthogonality.

In the present conditional-expectation setting, these identities provide the main estimator studied in this paper. Specifically, after flexibly estimating $\mathbb{E}[T \mid X, Z]$, $\mathbb{E}[T \mid X]$, $\mathbb{E}[Y \mid X, Z]$, and $\mathbb{E}[Y \mid X]$ on auxiliary folds, I form the examiner-specific representers as $\alpha_1(X, Z; \tilde{\theta}_l) = \mathbb{E}[Y \mid X, Z] - \tilde{\theta}_l \mathbb{E}[T \mid X, Z]$ and $\alpha_2(X; \tilde{\theta}_l) = -\mathbb{E}[Y \mid X] + \tilde{\theta}_l \mathbb{E}[T \mid X]$, where $\tilde{\theta}_l$ is a fold-specific pilot estimated on an additional auxiliary split. Thus the estimator combines an examiner-specific orthogonal score with explicit representers, evaluated at a cross-fitted pilot. Orthogonality ensures that this preliminary choice is asymptotically second order. The automatic-Riesz regression approach described below should therefore be interpreted as an extension of the central estimator for the examiner-IV problem studied here.

For the LASSO estimator in equation (3.13), researchers have to "choose" the penalty term, r, and the dictionary of functions, $b(X, Z)$. The penalty term, r, can be computed via a cross-validation procedure described in Appendix A.1.1 of Chernozhukov, Newey, and Singh (2022a), but other choices are also possible (see some examples in Chernozhukov, Escanciano, et al. (2022)). In practice, $b(X, Z)$ could be a fully-interacted specification with all covariates and examiner fixed effects or types, polynomials of continuous covariates, interactions of all covariates and examiner types, and interactions among covariates, as in Chernozhukov, Newey, and Singh (2022b). I emphasize, however, that this sparse-regression approach is not required for the main examiner-specific estimator, which instead uses the explicit representer identities above.

Estimation of $\alpha_2(X)$. The estimation of $\alpha_2(X)$ proceeds analogously. Holding γ_1 fixed and varying γ_2, the sample Gateaux derivative yields the moment condition $\mathbb{E}[(-Y + \theta T - \alpha_2)\delta] =$

0 for all $\delta \in L^2(X)$. Representing $\alpha_2(X) = \rho_2' b(X)$ where $b(X) = (b_1(X), \ldots, b_q(X))'$ is a dictionary of functions of X only, the LASSO estimator is

$$\hat{\rho}_2 = \underset{\rho_2}{\operatorname{argmin}} \frac{1}{n - n_l} \sum_{i \notin I_l} \Big((-Y_i + \tilde{\theta}_l T_i) - \rho_2' b(X_i) \Big)^2 + 2r_2 \sum_{j=1}^{q} |\rho_{2j}|,$$

and $\hat{\alpha}_{2l}(x) = \hat{\rho}_2' b(x)$. The key difference from the α_1 estimation is that the dictionary $b(X)$ is a function of covariates X only, excluding the examiner indicators Z.

Instead of the LASSO estimator, one could also use a Generalized Dantzig Selector (GDS) suggested in Chernozhukov, Newey, and Singh (2022b). Other approaches for learning the Riesz representers from data involve minimizing a stochastic loss over more general function spaces in which the true Riesz representer is hypothesized to live. Leveraging critical radius theory, Chernozhukov, Newey, Singh, and Syrgkanis (2020) provide an adversarial estimator of the Riesz representer over more general function spaces such as neural networks, random forests and reproducing kernel Hilbert spaces (in particular, the neural tangent kernel space). These more general approaches are computationally harder, and they are not needed for the main examiner-specific implementation studied here. I therefore present the LASSO formulation only as a simple illustrative example of the broader automatic-Riesz approach.

3.4 Asymptotic Theory

The large-sample theory for the method-of-moments estimator proposed in subsection 3.3.1 requires a small set of high-level conditions. First, the debiased sample moment must admit an asymptotically linear representation with influence function $\psi(W, \theta_0, \gamma_0, \alpha_0)$. To establish this, I impose mean-square consistency of the nuisance estimates together with a product-rate restriction on the interaction term. Second, I use this representation to derive asymptotic normality of the estimator. The discussion here covers the main examiner-specific implementation based on flexible estimation of the conditional means together with the explicit representer identities above. Generic automatic-Riesz estimators can be analyzed using the results already provided in Chernozhukov, Newey, and Singh (2022a), but that broader route is not needed for the main examiner-IV implementation.

Assumption 3.1 (Mean Square Consistency)**.** *(i) For each $i \in \{1, 2\}$, $\hat{\gamma}_{il}$ is a consistent estimator of γ_{0i} in the mean square sense such that $\|\hat{\gamma}_{il} - \gamma_{0i}\| \xrightarrow{p} 0$; (ii) For each $i \in \{1, 2\}$, $\hat{\alpha}_{il}(\theta_0)$ is a consistent estimator of $\alpha_{0i}(\theta_0)$ in the mean square sense such that $\|\hat{\alpha}_{il}(\theta_0) - \alpha_{0i}(\theta_0)\| \xrightarrow{p} 0$, where, in both (i) and (ii), $\|a\| = \sqrt{\mathbb{E}[a(W)^2]}$ is the L^2 norm.*

Assumption 3.2 (Pilot consistency, uniform convergence, and convergence rates for first steps)**.** *(i) $\theta_0 \in \Theta^o$, where Θ^o is the interior of the parameter space, Θ. The fold-specific*

pilot satisfies $\tilde{\theta}_l \xrightarrow{p} \theta_0$*, and for each* $i \in \{1,2\}$*,* $\hat{\alpha}_{il}(\theta)$ *converges uniformly on* Θ^o*. (ii) The nuisance estimates satisfy the product-rate condition*

$$\left\|\hat{\alpha}_l\left(\tilde{\theta}_l\right) - \alpha_0\left(\theta_0\right)\right\| \|\hat{\gamma}_l - \gamma_0\| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

A sufficient condition is that both factors are $o_p(n^{-1/4})$*, or more generally that one nuisance component converges fast enough to offset slower convergence of the other.*

Theorem 3.1 ($\sqrt{n}$ convergence of the debiased sample moments)**.** *Consider the estimator defined by (3.11), where the representers are evaluated at the fold-specific pilot* $\tilde{\theta}_l$*. Under assumptions 3.1 and 3.2,*

$$\sqrt{n}\hat{\psi}\left(\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right) + o_p(1).$$

Theorem 3.1 is an asymptotic orthogonality result. In essence, it asserts that the estimation of $\hat{\gamma}$ does not have an asymptotic effect on the orthogonal moment function. The theorem also reaffirms the validity of the orthogonal moment function as an influence function, and clarifies the conditions under which this holds. This follows from the fact that any asymptotically linear and locally regular estimator can be represented as an average of an influence function (Kosorok, 2008). For simplicity, I do not distinguish among the asymptotically equivalent preliminary estimators of θ used in constructing the nuisance functions; the fold-specific pilot $\tilde{\theta}_l$ is one such preliminary estimator.

Γ_1 and Γ_2 encode unrestricted square-integrable nuisance classes, with $\Gamma_1 = L^2(X, Z)$ and $\Gamma_2 = L^2(X)$. In that sense, the score is derived in a large nonparametric model rather than under additional shape restrictions. I do not pursue a separate semiparametric efficiency analysis here; for present purposes, Theorem 3.1 and Corollary 3.1 provide the asymptotic linearity and normality results needed for inference.

To prove Theorem 3.1, I invoke Lemma 8 in Chernozhukov, Escanciano, et al. (2022), which holds under the relevant mean-square consistency conditions. The proof therefore consists of verifying those conditions for the nonparametric first steps considered here. The details are provided in Appendix B.

Theorem 3.1 yields an important corollary: the asymptotic normality of the estimator. This corollary forms the basis for inference. We, however, need two additional assumptions.

Let $\sigma_{01}^2(X, Z) = \mathbb{E}\left[(T - \gamma_{01}(X, Z))^2 \mid X, Z\right]$ and $\sigma_{02}^2(X) = \mathbb{E}\left[(T - \gamma_{02}(X))^2 \mid X\right]$.

Assumption 3.3 (Boundedness)**.** *(i) For* $i \in \{1,2\}$*,* α_{0i} *and* α_{il} *are bounded; (ii)* $\sigma_{01}^2(X, Z)$ *and* $\sigma_{02}^2(X)$ *are bounded, and*

$$\mathbb{E}[g(W, \gamma_0, \theta_0)^2] < \infty.$$

Because the debiased sample moment in (3.11) is affine in θ once the cross-fitted nuisance estimates and fold-specific pilot values are held fixed, its derivative with respect to θ is constant:

$$\widehat{Q} = \frac{\partial \hat{\psi}(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} T_i \left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i) \right).$$

For asymptotic normality, it is therefore enough to assume that this sample Jacobian converges to a nonzero population limit.

Assumption 3.4 (Nondegenerate Jacobian)**.**

$$Q := -\mathbb{E}\left[T \left(\gamma_{01}(X, Z) - \gamma_{02}(X) \right) \right] \neq 0$$

and

$$\widehat{Q} \xrightarrow{p} Q.$$

Corollary 3.1 (Asymptotic Normality of the Method of Moments Estimator)**.** *Under Assumptions 3.1–3.4, the MoM estimator is asymptotically normal:*

$$\sqrt{n} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} N(0, V) \quad \text{and} \quad \hat{V} \xrightarrow{p} V$$

where $V = Q^{-2} \operatorname{Var}(\psi) = Q^{-2} \mathbb{E}[\psi^2]$ *because* $\mathbb{E}[\psi] = 0$*, and* $\hat{V}$ *is a consistent estimator of* V*.*

Appendix B contains the proof of this corollary. Heuristically, the variance expression follows from the asymptotically linear representation in Theorem 3.1 together with standard arguments for Z-estimators. The result provides the basis for standard confidence intervals and hypothesis tests.

Corollary 3.1 presumes $\sqrt{n}$ consistency of the estimator, $\hat{\theta}$. Primitive conditions for this consistency are mild and are provided in Appendix F of the Supplemental Material in Chernozhukov, Escanciano, et al. (2022). In a nutshell, consistency of the estimator requires, *inter alia*, the mean square consistency conditions of Assumption 1 in Chernozhukov, Escanciano, et al. (2022), which I have proved for Theorem 3.1 in our setting; the rate condition in Assumption 2 in Chernozhukov, Escanciano, et al. (2022), which I have also proved for Theorem 3.1; and standard consistency assumptions, such as compactness of the parameter space, from Newey and McFadden (1994).

4 Monte Carlo Evidence

This section provides Monte Carlo evidence on the finite-sample behavior of the orthogonal estimator proposed in this paper. I compare three estimators throughout: the examiner-

specific orthogonal estimator, a naive flexible plug-in estimator that uses the same first-step regressions but omits the orthogonal correction, and the linear UJIVE of Kolesár (2013). The goal of the exercise is to illustrate the practical implications of the orthogonal correction under a transparent sieve-based implementation that mirrors the estimator studied in Section 3, rather than to survey every available first-step learner. All first-step conditional expectation functions are estimated by cross-fitted ridge-regularized sieve regressions, and the orthogonal estimator is constructed with the nested cross-fitted pilot procedure described in Section 3.3.1.

I consider two data-generating processes that serve distinct purposes. The first is a benchmark design in which the implemented sieve is well aligned with the true conditional mean functions, so that the conditions of Theorem 3.1 are approximately satisfied. The second is a nonlinear design in which the treatment propensity is intentionally placed outside the sieve span. This separation allows one to distinguish evidence on the finite-sample behavior of asymptotic inference under the maintained assumptions from evidence on point-estimation performance under first-step misspecification. To keep the comparison transparent, both designs use the same sample size and the same number of examiners.

Design 1 (benchmark). The sample consists of $n = 1{,}800$ observations randomly assigned to $J = 18$ examiners with equal probability. Covariates $X = (X_1, \dots, X_6)$ are drawn independently from the uniform distribution on $[-1, 1]$. Examiner leniency parameters ℓ_j are equally spaced on $[-1.25, 1.25]$. Treatment is binary with

$$T \mid X, Z = j \sim \text{Bernoulli}(p(X, j)),$$

where

$$p(X, j) = 0.50 + 0.05\, X_1 + 0.30\, \ell_j.$$

The treatment propensity therefore depends on both covariates and examiner assignment while remaining well approximated by the implemented first-step sieve, so that the nuisance estimation conditions entering Assumptions 3.1 and 3.2 are approximately in force. Outcomes satisfy

$$Y = \mu(X) + \tau(X)T + 0.20\, \varepsilon, \qquad \varepsilon \sim N(0, 1),$$

with heterogeneous treatment effects $\tau(X) = 1 + 0.30\, X_1$ and a nonlinear baseline regression

$$\mu(X) = 0.6 + 0.50 \sin(0.8\, X_1) - 0.30\, X_2 + 0.35\, X_3 X_4 - 0.30\, X_5^2 + 0.25\, \mathbf{1}\{X_2 > 0\} + 0.20\, X_6^2.$$

The baseline function μ is nonlinear but lies in the span of the sieve dictionary used to estimate the outcome regressions. This design is therefore intended to assess the finite-sample performance of asymptotic inference under conditions favorable to the estimator.

Design 2 (nonlinear). The sample consists of $n = 1{,}800$ observations randomly assigned to $J = 18$ examiners with equal probability. Covariates $X = (X_1, \dots, X_6)$ have independent standard normal coordinates. Examiner leniency parameters are equally spaced on $[-1.5625, 1.5625]$. Treatment is binary with

$$T \mid X, Z = j \sim \text{Bernoulli}\Big(\Lambda\big(\ell_j\big(1 + 0.95\,X_1 - 0.75\,\mathbf{1}\{X_2 > 0\}\big) + 1.10\sin(0.8\,X_1) - 0.25\,X_2 + 0.50\,X_3X_4 - 0.40\,X_5^2 + 0.25\,X_6\big)\Big),$$

where $\Lambda(u) = 1/(1 + e^{-u})$. The judge-by-covariate interactions inside the logistic link, together with the nonlinear covariate terms, place the treatment propensity outside the span of the implemented sieve. Outcomes satisfy

$$Y = \mu(X) + \tau(X)\,T + 0.25\,\varepsilon, \qquad \varepsilon \sim N(0, 1),$$

with $\mu(X) = 0.5 + 0.35\,X_1 - 0.20\,X_2 + 0.25\,X_3X_4 - 0.25\,X_5 + 0.15\,X_6^2$ and $\tau(X) = 1 + 0.40\,X_1 + 0.25\,\mathbf{1}\{X_2 > 0\}$. Because the first step is misspecified by construction, this design should be read as a stress test of point-estimation performance rather than as a benchmark for valid asymptotic inference.

Table 1 reports the results. For the benchmark design I report Monte Carlo mean, bias, RMSE, standard deviation, nominal 95% coverage, and the average reported standard error based on 100 replications. For the nonlinear design I report only mean, bias, RMSE, and standard deviation based on 200 replications, since that design is not intended to assess inferential calibration and reporting coverage there would invite a misleading comparison. In both designs the true parameter θ_0 is approximated by a large auxiliary simulation under the same data-generating process.

Table 1: Monte Carlo results: benchmark and nonlinear designs

Design	Estimator	Mean	Bias	RMSE	SD	Coverage	Avg. SE
Benchmark	Orthogonal	0.997	−0.003	0.033	0.033	0.930	0.030
Benchmark	Plug-in	0.999	−0.002	0.043	0.043	—	—
Benchmark	Linear UJIVE	0.998	−0.003	0.029	0.029	1.000	0.085
Nonlinear	Orthogonal	1.432	0.058	0.090	0.069	—	—
Nonlinear	Plug-in	1.260	−0.114	0.151	0.099	—	—
Nonlinear	Linear UJIVE	1.587	0.213	0.241	0.114	—	—

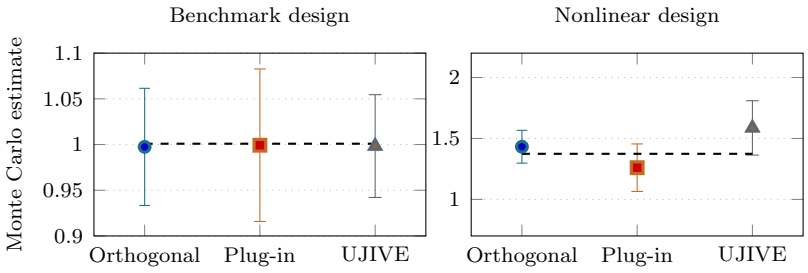

Figure 1: Monte Carlo means with 95% simulation intervals. Whiskers equal ±1.96 empirical standard deviations across replications; dashed lines mark the true parameter values.

Figure 1 provides a graphical summary of the same Monte Carlo evidence. Appendix Figure 2 complements the table with RMSE and uncertainty comparisons.

Consider first the benchmark design. The orthogonal estimator is essentially unbiased (bias $= -0.003$) and attains substantially lower RMSE than the naive plug-in estimator (0.033 versus 0.043) while using the same cross-fitted first-step regressions. Coverage is 93%, close to the nominal 95% level, though the average reported standard error (0.030) remains slightly below the empirical standard deviation (0.033). This modest undercoverage is consistent with findings in the broader double machine learning literature, where finite-sample inference is known to remain sensitive to the quality of first-step estimation and to the particular realization of the sample split, especially at moderate sample sizes; see, for example, Zivich and Breskin (2021) and Ahrens et al. (2024). Linear UJIVE is also nearly unbiased in this design and attains lower RMSE (0.029), reflecting the fact that the linear examiner effect is correctly specified by construction so that the more parsimonious linear basis yields sharper estimates. Its reported standard errors (0.085) are markedly conservative relative to the empirical standard deviation (0.029), producing 100% coverage; this conservativeness comes from the variance estimator in this design rather than from the point estimator itself. The benchmark design is therefore not intended as a horse race against UJIVE on point-estimation efficiency; rather, it is intended to verify that the orthogonal estimator delivers coverage close to nominal when the maintained assumptions are approximately in force.

The nonlinear design serves a different purpose. Because the treatment propensity lies outside the implemented sieve, the table should be read primarily as evidence on finite-sample bias and RMSE under first-step misspecification. For that reason, I do not report coverage or average standard errors for this design. Here the orthogonal estimator improves substantially on both competitors: relative to linear UJIVE, bias falls from 0.213 to 0.058 and RMSE from 0.241 to 0.090; relative to the naive plug-in, RMSE falls from 0.151 to 0.090 while using the same flexible first-step regressions. This pattern is consistent with the orthogonal correction absorbing a substantial share of the first-step estimation error, even

when the product-rate condition in Assumption 3.2 is not fully satisfied. At the same time, the residual bias (0.058) confirms that orthogonalization does not eliminate the dependence of the estimator on first-step quality when the propensity is genuinely outside the sieve span.

Taken together, the two designs illustrate three points. First, when the first-step sieve is well aligned with the data-generating process, the orthogonal estimator is essentially unbiased and achieves coverage close to nominal. Second, in harder nonlinear environments where the first step is misspecified, orthogonalization yields large improvements in bias and RMSE relative to both the linear benchmark and the naive plug-in estimator. Third, valid asymptotic inference ultimately depends on the quality of the first-step approximation; the orthogonal correction mitigates but does not eliminate this dependence.

5 Conclusion

This paper has provided an examiner-specific semiparametric framework for estimation and asymptotic inference for the covariate-adjusted Kolesár (2013) examiner-IV estimand when the generated leniency instrument must be estimated flexibly. The main estimator combines the identifying ratio moment for this standard examiner-IV target with an orthogonal influence-function adjustment and explicit representer identities tailored to the examiner setting. Based on the locally robust semiparametric theory of Chernozhukov, Escanciano, et al. (2022) and Ichimura and Newey (2022), I provide conditions under which this two-step estimator remains root-n consistent for first-step estimators that satisfy the stated mean-square consistency and product-rate conditions.

As an avenue of future research, it would also be interesting to explore locally robust semiparametric approaches for settings where examiners administer more than one treatment, such as in Kamat et al. (2023). Furthermore, the method developed in this paper presumes that the canonical identification assumptions of these examiner IV designs obtain, but it might be instructive to extend this framework to settings where some of these assumptions do not hold, a la Frandsen, Lefgren, et al. (2023).

A Additional Monte Carlo Figure

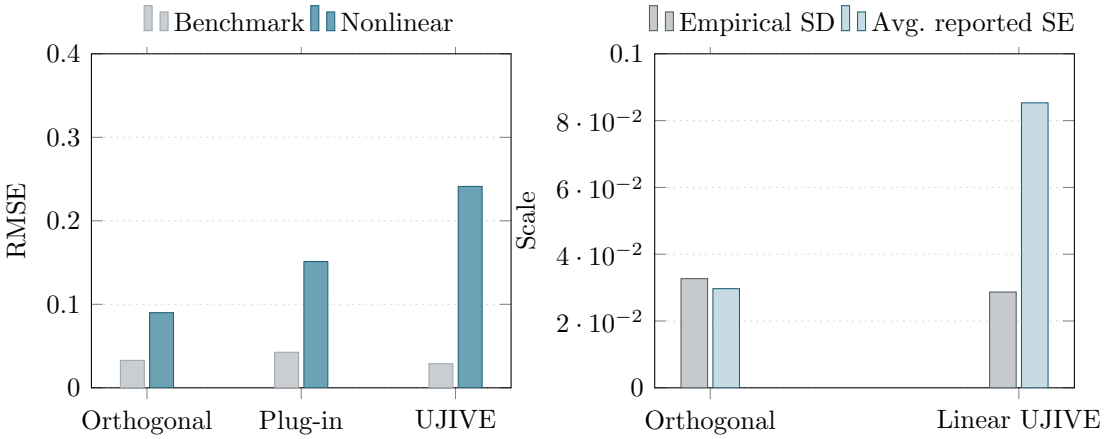

Figure 2: Simulation performance and benchmark inference summary. Left: RMSE by estimator and design. Right: in the benchmark design, the orthogonal estimator's average reported standard error is close to the empirical standard deviation, whereas linear UJIVE is markedly conservative.

B Proofs of the Asymptotic Theory

Lemma B.1 (Exact bilinear decomposition of the orthogonal score)**.** *Let*

$$\mu_0(W) = Y - \theta_0 T, \qquad \gamma_0 = (\gamma_{01}, \gamma_{02}), \qquad \alpha_0 = (\alpha_{01}, \alpha_{02}),$$

where

$$\gamma_{01}(X, Z) = \mathbb{E}[T \mid X, Z], \qquad \gamma_{02}(X) = \mathbb{E}[T \mid X],$$

and

$$\alpha_{01}(X, Z) = \mathbb{E}[\mu_0(W) \mid X, Z], \qquad \alpha_{02}(X) = -\mathbb{E}[\mu_0(W) \mid X].$$

Then for any square-integrable nuisance functions

$$\gamma = (\gamma_1, \gamma_2), \qquad \alpha = (\alpha_1, \alpha_2),$$

the orthogonal score in (3.7) *satisfies*

$$\mathbb{E}\big[\psi(W, \gamma, \theta_0, \alpha)\big] = -\mathbb{E}\Big[(\alpha_1 - \alpha_{01})(\gamma_1 - \gamma_{01})\Big] - \mathbb{E}\Big[(\alpha_2 - \alpha_{02})(\gamma_2 - \gamma_{02})\Big].$$

In particular, the score remains valid whenever, for each $j \in \{1, 2\}$, at least one of α_j or γ_j is correctly specified.

Proof. Write

$$\delta\gamma_1 = \gamma_1 - \gamma_{01}, \qquad \delta\gamma_2 = \gamma_2 - \gamma_{02}, \qquad \delta\alpha_1 = \alpha_1 - \alpha_{01}, \qquad \delta\alpha_2 = \alpha_2 - \alpha_{02}.$$

Since $\mathbb{E}[\psi(W, \gamma_0, \theta_0, \alpha_0)] = 0$, it is enough to expand

$$\psi(W, \gamma, \theta_0, \alpha) - \psi(W, \gamma_0, \theta_0, \alpha_0).$$

Using $\mu_0(W) = Y - \theta_0 T$, a direct calculation gives

$$\begin{aligned}
&\psi(W, \gamma, \theta_0, \alpha) - \psi(W, \gamma_0, \theta_0, \alpha_0) \\
&\quad = (\mu_0 - \alpha_{01})\delta\gamma_1 + (-\mu_0 - \alpha_{02})\delta\gamma_2 + \delta\alpha_1(T - \gamma_{01}) + \delta\alpha_2(T - \gamma_{02}) - \delta\alpha_1\delta\gamma_1 - \delta\alpha_2\delta\gamma_2.
\end{aligned}$$

Taking expectations, the first four terms vanish. Indeed,

$$\mathbb{E}\big[(\mu_0 - \alpha_{01})h_1(X, Z)\big] = 0 \quad \text{for all } h_1 \in L^2(X, Z),$$

because $\alpha_{01}(X,Z) = \mathbb{E}[\mu_0 \mid X, Z]$, and

$$\mathbb{E}\big[(-\mu_0 - \alpha_{02})h_2(X)\big] = 0 \quad \text{for all } h_2 \in L^2(X),$$

because $\alpha_{02}(X) = -\mathbb{E}[\mu_0 \mid X]$. Likewise,

$$\mathbb{E}\big[a_1(X,Z)(T - \gamma_{01})\big] = 0 \quad \text{for all } a_1 \in L^2(X,Z),$$

because $\gamma_{01}(X,Z) = \mathbb{E}[T \mid X, Z]$, and

$$\mathbb{E}\big[a_2(X)(T - \gamma_{02})\big] = 0 \quad \text{for all } a_2 \in L^2(X),$$

because $\gamma_{02}(X) = \mathbb{E}[T \mid X]$. Therefore

$$\mathbb{E}\big[\psi(W, \gamma, \theta_0, \alpha)\big] = -\mathbb{E}\big[\delta\alpha_1\delta\gamma_1\big] - \mathbb{E}\big[\delta\alpha_2\delta\gamma_2\big],$$

which is the claimed identity. □

Corollary B.1 (Exact product-rate bound)**.** *For any fold l,*

$$\big|\mathbb{E}\big[\psi(W, \hat{\gamma}_l, \theta_0, \hat{\alpha}_l)\big]\big| \le \|\hat{\alpha}_{1l} - \alpha_{01}\|_2 \, \|\hat{\gamma}_{1l} - \gamma_{01}\|_2 + \|\hat{\alpha}_{2l} - \alpha_{02}\|_2 \, \|\hat{\gamma}_{2l} - \gamma_{02}\|_2.$$

Hence the bias from first-step estimation enters through a bilinear interaction term.

Proof. This follows immediately from Lemma B.1 and Cauchy–Schwarz. □

Remark B.1 (What is exact here)**.** *Neyman orthogonality remains a local derivative property. What is exact in the present score is the bilinear decomposition in Lemma B.1. In particular, along any path*

$$\gamma_t = \gamma_0 + t h_\gamma, \qquad \alpha_t = \alpha_0 + t h_\alpha,$$

we have

$$\mathbb{E}\big[\psi(W, \gamma_t, \theta_0, \alpha_t)\big] = -t^2\mathbb{E}[h_{\alpha_1} h_{\gamma_1}] - t^2\mathbb{E}[h_{\alpha_2} h_{\gamma_2}],$$

so the Gateaux derivative at $t = 0$ is zero because the population moment error contains no linear term.

B.1 Proof of Theorem 3.1

I verify the mean-square consistency conditions of Assumption 3.1 and the product-rate condition in Assumption 3.2 for Lemma 8 in Section 6 of Chernozhukov, Escanciano, et al.

(2022). Throughout, C denotes a generic finite constant whose value may change from line to line.

Assumption 1(i) of Lemma 8 is:

$$\int \|g(w, \hat{\gamma}_\ell, \theta_0) - g(w, \gamma_0, \theta_0)\|^2 F_0(dw) \xrightarrow{p} 0$$

For notational economy, I treat $\hat{\gamma}_l$ as a single first-step object rather than explicitly writing it as a difference of two functions:

$$g(W, \hat{\gamma}_l, \theta_0) - g(W, \gamma_0, \theta_0) = (Y - \theta_0 T)(\hat{\gamma}_l - \gamma_0).$$

Specialized to $g(W, \gamma, \theta) = (Y - \theta T)\gamma$, Assumption 1(i) of Lemma 8 in Chernozhukov, Escanciano, et al. (2022) is continuity of the map $\gamma \mapsto g(W, \gamma, \theta_0)$ in mean square at γ_0. A sufficient primitive condition is $\mathbb{E}[(Y - \theta_0 T)^4] < \infty$ together with $\|\hat{\gamma}_l - \gamma_0\|_4 \xrightarrow{p} 0$, where $\|\cdot\|_4$ denotes the L^4 norm. In that case, Cauchy–Schwarz gives

$$\int \|g(w, \hat{\gamma}_\ell, \theta_0) - g(w, \gamma_0, \theta_0)\|^2 F_0(dw) \leq \mathbb{E}\big[(Y - \theta_0 T)^4\big]^{1/2} \mathbb{E}\big[(\hat{\gamma}_l - \gamma_0)^4\big]^{1/2} \xrightarrow{p} 0.$$

I proceed under the corresponding high-level mean-square continuity condition implied by Assumptions 3.1 and 3.3. Therefore,

$$\int \|g(w, \hat{\gamma}_\ell, \theta_0) - g(w, \gamma_0, \theta_0)\|^2 F_0(dw) \xrightarrow{p} 0,$$

which is the desired claim.

Assumption (ii) of Lemma 8 in Chernozhukov, Escanciano, et al. (2022) imposes the following restriction:

$$\int \|\phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) - \phi(w, \gamma_0, \alpha_0, \theta_0)\|^2 F_0(dw) \xrightarrow{p} 0$$

In our setting,

$$\phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) = \alpha_{01}(T - \hat{\gamma}_{1\ell}) + \alpha_{02}(T - \hat{\gamma}_{2\ell})$$

and

$$\phi(w, \gamma_0, \alpha_0, \theta_0) = \alpha_{01}(T - \gamma_{01}) + \alpha_{02}(T - \gamma_{02}).$$

Then

$$\begin{aligned}
&\|\phi\left(w, \hat{\gamma}_{\ell}, \alpha_0, \theta_0\right) - \phi\left(w, \gamma_0, \alpha_0, \theta_0\right)\|^2 \\
&= \|\alpha_{01}\left(\gamma_{01} - \hat{\gamma}_{1\ell}\right) + \alpha_{02}\left(\gamma_{02} - \hat{\gamma}_{2\ell}\right)\|^2 \\
&\leq 2\|\alpha_{01}\left(\gamma_{01} - \hat{\gamma}_{1\ell}\right)\|^2 + 2\|\alpha_{02}\left(\gamma_{02} - \hat{\gamma}_{2\ell}\right)\|^2 \\
&\leq 2C_1^2\|\gamma_{01} - \hat{\gamma}_{1\ell}\|^2 + 2C_2^2\|\hat{\gamma}_{2\ell} - \gamma_{02}\|^2.
\end{aligned}$$

The penultimate inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$, and the last inequality follows from boundedness of the Riesz representers.

Then it follows that:

$$\int \|\phi\left(w, \hat{\gamma}_{\ell}, \alpha_0, \theta_0\right) - \phi\left(w, \gamma_0, \alpha_0, \theta_0\right)\|^2 F_0(dw) \leq 2C_1^2\|\gamma_{01} - \hat{\gamma}_{1\ell}\|^2 + 2C_2^2\|\hat{\gamma}_{2\ell} - \gamma_{02}\|^2.$$

The result follows by Assumption 3.1.

Assumption 1(iii) in Chernozhukov, Escanciano, et al. (2022) requires:

$$\int \left\|\phi\left(w, \gamma_0, \hat{\alpha}_l, \tilde{\theta}_l\right) - \phi\left(w, \gamma_0, \alpha_0, \theta_0\right)\right\|^2 F_0(dw) \xrightarrow{p} 0$$

Suppressing the dependence of the Riesz representers and their estimators on X and Z for notational convenience and letting them depend on only θ_0 and $\tilde{\theta}_l$ respectively:

$$\begin{aligned}
&\int \left\|\phi(w, \gamma_0, \hat{\alpha}_l, \tilde{\theta}_l) - \phi(w, \gamma_0, \alpha_0, \theta_0)\right\|^2 F_0(dw) \\
&= \int \left\|\hat{\alpha}_{1l}(\tilde{\theta}_l)(T - \gamma_{01}) + \hat{\alpha}_{2l}(\tilde{\theta}_l)(T - \gamma_{02}) - \alpha_{01}(\theta_0)(T - \gamma_{01}) - \alpha_{02}(\theta_0)(T - \gamma_{02})\right\|^2 F_0(dw) \\
&\leq 2\int \left\|(\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0))(T - \gamma_{01})\right\|^2 F_0(dw) \\
&\quad + 2\int \left\|\left(\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)\right)(T - \gamma_{02})\right\|^2 F_0(dw) \\
&= \int (T - \gamma_{01})^2 \left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\right\|^2 F_0(dw) + \int (T - \gamma_{02})^2 \left\|\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)\right\|^2 F_0(dw) \\
&\leq C_1 \left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\right\|^2 + C_2 \left\|\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)\right\|^2.
\end{aligned}$$

Take $C_1 \left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\right\|^2$ and observe that the problem of $C_2 \left\|\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)\right\|^2$ is

completely symmetric:

$$\left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\right\|^2 = \left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \hat{\alpha}_{1l}(\theta_0) + \hat{\alpha}_{1l}(\theta_0) - \alpha_{01}(\theta_0)\right\|^2$$
$$\leq 2\left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \hat{\alpha}_{1l}(\theta_0)\right\|^2 + 2\left\|\hat{\alpha}_{1l}(\theta_0) - \alpha_{01}(\theta_0)\right\|^2.$$

where the last inequality follows from $(a+b)^2 \leq 2a^2 + 2b^2$. By Assumption 3.2, $\tilde{\theta}_l$ converges in probability to θ_0, and by the continuous mapping theorem, $\hat{\alpha}_{1l}(\tilde{\theta}_l)$ also converges in probability to $\hat{\alpha}_{1l}(\theta_0)$. It follows that $\left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \hat{\alpha}_{1l}(\theta_0)\right\|^2$ goes to zero. By the uniform convergence property in Assumption 3.2, we know that $\hat{\alpha}_{1l}(\theta_0)$ converges to $\alpha_{01}(\theta_0)$. Therefore, $C_1\left\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\right\|^2$ converges to zero and, by symmetry, $C_2\left\|\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)\right\|^2$ also converges to zero. It follows, therefore, that as desired:

$$\int \left\|\phi\left(w, \gamma_0, \hat{\alpha}_l, \tilde{\theta}_l\right) - \phi\left(w, \gamma_0, \alpha_0, \theta_0\right)\right\|^2 F_0(dw) \xrightarrow{p} 0$$

Let $\hat{\Delta}_l(w) = \phi\left(w, \hat{\gamma}_l, \hat{\alpha}_l, \tilde{\theta}_l\right) - \phi\left(w, \gamma_0, \hat{\alpha}_l, \tilde{\theta}_l\right) - \phi\left(w, \hat{\gamma}_l, \alpha_0, \theta_0\right) + \phi\left(w, \gamma_0, \alpha_0, \theta_0\right)$. This is what Chernozhukov, Escanciano, et al. (2022) call "an interaction term". Chernozhukov, Escanciano, et al. (2022) impose a rate condition on this interaction term. I prove that the rate condition in Assumption 2(i) in Chernozhukov, Escanciano, et al. (2022) holds in this setting, namely:

$$\sqrt{n}\int \hat{\Delta}_l(w)F_0(dw) \xrightarrow{p} 0, \quad \int \left\|\hat{\Delta}_\ell(w)\right\|^2 F_0(dw) \xrightarrow{p} 0$$

Treating γ as one first step function rather than a difference of two functions to avoid notational clutter:

$$\int \hat{\Delta}_\ell(w)F_0(dw)$$
$$= \int (\hat{\alpha}_l(\tilde{\theta}_l)(T - \hat{\gamma}_l) - \hat{\alpha}_l(\tilde{\theta}_l)(T - \gamma_0)) - \alpha_0(\theta_0)(T - \hat{\gamma}_l) + \alpha_0(\theta_0)(T - \gamma_0)F_0(dw)$$
$$= \int (\hat{\alpha}_l(\tilde{\theta}_l) - \alpha_0(\theta_0))(\gamma_0 - \hat{\gamma}_l)F_0(dw) \leq \left\|\hat{\alpha}_l(\tilde{\theta}_l) - \alpha_0(\theta_0)\right\| \|\hat{\gamma}_l - \gamma_0\|$$
$$= o_p\left(\frac{1}{\sqrt{n}}\right).$$

The inequality follows from applying the Cauchy-Schwarz inequality, and the last equality follows from Assumption 3.2(ii). This proves $\sqrt{n}\int \hat{\Delta}_\ell(w)F_0(dw) \xrightarrow{p} 0$.

We now need to show $\int \left\|\hat{\Delta}_\ell(w)\right\|^2 F_0(dw) \xrightarrow{p} 0$. This follows from the same product-rate argument:

$$\int \|\hat{\Delta}_l(w)\|^2 F_0(dw) \leq \left\|\hat{\alpha}_l\left(\tilde{\theta}_l\right) - \alpha_0\left(\theta_0\right)\right\|^2 \|\hat{\gamma}_l - \gamma_0\|^2 = o_p(1)$$

where the last equality follows from Assumption 3.2(ii).

Assumption 3(i) and (ii) of Lemma 8 in Chernozhukov, Escanciano, et al. (2022) are also satisfied here. Assumption 3(i) follows from consistency of $\tilde{\theta}_l$ for θ_0, the convergence of $\hat{\alpha}_l(\tilde{\theta}_l)$ to $\alpha_0(\theta_0)$ established above, and the fact that the first-step influence function has mean zero at the true nuisance values. Assumption 3(ii) requires the orthogonal moment function to be affine in the first-step nuisance functions, which has already been verified in Section 3.2.

Having verified the three conditions of Lemma 8 in Chernozhukov, Escanciano, et al. (2022), its conclusion follows immediately:

$$\sqrt{n}\hat{\psi}\left(\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right) + o_p(1)$$

B.2 Proof of Corollary 3.1

The proof proceeds by combining Theorem 3.1 with the fact that the sample moment is affine in θ.

Since $\hat{\psi}(\hat{\theta}) = 0$ and $\hat{\psi}(\theta)$ is linear in θ, we have the exact identity

$$0 = \hat{\psi}(\theta_0) + \widehat{Q}(\hat{\theta} - \theta_0),$$

where

$$\widehat{Q} = \frac{\partial \hat{\psi}(\theta)}{\partial \theta} = -\frac{1}{n}\sum_{l=1}^{L}\sum_{i \in I_l} T_i\left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i)\right).$$

Rearranging gives

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\widehat{Q}^{-1}\sqrt{n}\hat{\psi}(\theta_0).$$

By Assumption 3.4, $\widehat{Q}^{-1} = Q^{-1} + o_p(1)$, and by Theorem 3.1,

$$\sqrt{n}\hat{\psi}\left(\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right) + o_p(1).$$

Therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -Q^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right) + o_p(1) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = Q^{-2} \operatorname{Var}(\psi) = Q^{-2}\mathbb{E}[\psi^2]$ since $\mathbb{E}[\psi] = 0$.

The estimator for V is:

$$\hat{V} = \widehat{Q}^{-2} \frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{\psi}_i(\hat{\theta})^2,$$

where

$$\begin{aligned} \hat{\psi}_i(\theta) &= \left(Y_i - \theta T_i\right)\left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i)\right) \\ &\quad + \hat{\alpha}_{1l}(X_i, Z_i; \tilde{\theta}_l)\left(T_i - \hat{\gamma}_{1l}(X_i, Z_i)\right) \\ &\quad + \hat{\alpha}_{2l}(X_i; \tilde{\theta}_l)\left(T_i - \hat{\gamma}_{2l}(X_i)\right) \end{aligned}$$

for $i \in I_l$, and

$$\widehat{Q} = \frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \left.\frac{\partial \hat{\psi}_i(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = -\frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} T_i \left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i)\right).$$

To prove consistency of $\hat{V}$, first note that

$$\hat{\psi}_i(\theta_0) - \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right)$$

is a sum of terms involving $\hat{\gamma}_l - \gamma_0$ and $\hat{\alpha}_l(\tilde{\theta}_l) - \alpha_0(\theta_0)$. By the same decompositions used in the proof of Theorem 3.1, together with Assumptions 3.1–3.3,

$$\frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \left(\hat{\psi}_i(\theta_0) - \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right)\right)^2 \xrightarrow{p} 0.$$

Hence

$$\frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{\psi}_i(\theta_0)^2 = \frac{1}{n} \sum_{i=1}^{n} \psi\left(W_i, \theta_0, \gamma_0, \alpha_0\right)^2 + o_p(1) \xrightarrow{p} \mathbb{E}\left[\psi\left(W, \theta_0, \gamma_0, \alpha_0\right)^2\right].$$

Next, because the pilot-frozen sample moment is affine in θ,

$$\hat{\psi}_i(\hat{\theta}) - \hat{\psi}_i(\theta_0) = -(\hat{\theta} - \theta_0)\, T_i \left(\hat{\gamma}_{1l}(X_i, Z_i) - \hat{\gamma}_{2l}(X_i)\right)$$

for $i \in I_l$. Therefore,

$$\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left(\hat{\psi}_i(\hat{\theta})-\hat{\psi}_i(\theta_0)\right)^2 \le (\hat{\theta}-\theta_0)^2\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left(\hat{\gamma}_{1l}(X_i,Z_i)-\hat{\gamma}_{2l}(X_i)\right)^2 = o_p(1),$$

since $\hat{\theta}\xrightarrow{p}\theta_0$ and the sample average on the right is $O_p(1)$ by Assumptions 3.1 and 3.3. By Cauchy-Schwarz,

$$\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\hat{\psi}_i(\hat{\theta})^2 = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\hat{\psi}_i(\theta_0)^2 + o_p(1) \xrightarrow{p} \mathbb{E}\left[\psi\left(W,\theta_0,\gamma_0,\alpha_0\right)^2\right].$$

Combining this with Assumption 3.4 gives

$$\hat{V} = \widehat{Q}^{-2}\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\hat{\psi}_i(\hat{\theta})^2 \xrightarrow{p} Q^{-2}\mathbb{E}\left[\psi\left(W,\theta_0,\gamma_0,\alpha_0\right)^2\right] = V.$$

C Primitive Conditions for Assumptions 3.1 and 3.2

Assumptions 3.1 and 3.2 are stated in terms of high-level mean-square consistency and product-rate conditions. For the main examiner-specific implementation, however, the exact bilinear decomposition in Lemma B.1 together with the closed-form representers imply that these high-level assumptions can be reduced to low-level conditions on four conditional mean regressions. This appendix makes that reduction explicit and then records primitive examples.

Let

$$m_{01}(X,Z)=\mathbb{E}[Y\mid X,Z], \qquad m_{02}(X)=\mathbb{E}[Y\mid X],$$
$$g_{01}(X,Z)=\mathbb{E}[T\mid X,Z], \qquad g_{02}(X)=\mathbb{E}[T\mid X].$$

By the identities in Section 3.3,

$$\alpha_{01}(X,Z;\theta_0)=m_{01}(X,Z)-\theta_0 g_{01}(X,Z), \qquad \alpha_{02}(X;\theta_0)=-m_{02}(X)+\theta_0 g_{02}(X).$$

Accordingly, for each fold l I construct the closed-form nuisance estimates as

$$\hat{\alpha}_{1l}(x,z;\theta)=\hat{m}_{1l}(x,z)-\theta\hat{g}_{1l}(x,z), \qquad \hat{\alpha}_{2l}(x;\theta)=-\hat{m}_{2l}(x)+\theta\hat{g}_{2l}(x),$$

where $\hat{m}_{1l}$ and $\hat{m}_{2l}$ estimate $\mathbb{E}[Y\mid X,Z]$ and $\mathbb{E}[Y\mid X]$, while $\hat{g}_{1l}$ and $\hat{g}_{2l}$ estimate $\mathbb{E}[T\mid X,Z]$ and $\mathbb{E}[T\mid X]$.

Proposition C.1 (Reduction to conditional-mean rates)**.** *Suppose* Θ° *is compact,* $\|\hat{g}_{1l}\| + \|\hat{g}_{2l}\| = O_p(1)$*, and*

$$\|\hat{m}_{1l} - m_{01}\| + \|\hat{m}_{2l} - m_{02}\| + \|\hat{g}_{1l} - g_{01}\| + \|\hat{g}_{2l} - g_{02}\| \xrightarrow{p} 0.$$

Then Assumption 3.1 holds for the closed-form representer estimates above, and the uniform convergence part of Assumption 3.2(i) follows automatically. Moreover,

$$\left\|\hat{\alpha}_l(\tilde{\theta}_l) - \alpha_0(\theta_0)\right\| \lesssim \sum_{j=1}^{2} \left(\|\hat{m}_{jl} - m_{0j}\| + \|\hat{g}_{jl} - g_{0j}\| + |\tilde{\theta}_l - \theta_0|\right),$$

where m_{0j} *and* g_{0j} *denote the corresponding true conditional means. Hence Assumption 3.2(ii) is implied by*

$$\left(\sum_{j=1}^{2} \|\hat{m}_{jl} - m_{0j}\| + \|\hat{g}_{jl} - g_{0j}\| + |\tilde{\theta}_l - \theta_0|\right) (\|\hat{g}_{1l} - g_{01}\| + \|\hat{g}_{2l} - g_{02}\|) = o_p(n^{-1/2}).$$

In particular, if the pilot is root-n consistent and all four conditional-mean regressions converge at rate $o_p(n^{-1/4})$ *in* L^2*, then Assumptions 3.1 and 3.2 are satisfied.*

Proof. For the first representer,

$$\hat{\alpha}_{1l}(\theta) - \alpha_{01}(\theta) = (\hat{m}_{1l} - m_{01}) - \theta(\hat{g}_{1l} - g_{01}),$$

so compactness of Θ° gives

$$\sup_{\theta \in \Theta^{\circ}} \|\hat{\alpha}_{1l}(\theta) - \alpha_{01}(\theta)\| \leq \|\hat{m}_{1l} - m_{01}\| + \sup_{\theta \in \Theta^{\circ}} |\theta| \, \|\hat{g}_{1l} - g_{01}\| \xrightarrow{p} 0.$$

The same argument applies to $\hat{\alpha}_{2l}(\theta)$. This yields Assumption 3.1(ii) and the uniform convergence part of Assumption 3.2(i). Next,

$$\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0) = (\hat{m}_{1l} - m_{01}) - \theta_0(\hat{g}_{1l} - g_{01}) - (\tilde{\theta}_l - \theta_0)\hat{g}_{1l},$$

hence

$$\|\hat{\alpha}_{1l}(\tilde{\theta}_l) - \alpha_{01}(\theta_0)\| \leq \|\hat{m}_{1l} - m_{01}\| + |\theta_0|\|\hat{g}_{1l} - g_{01}\| + |\tilde{\theta}_l - \theta_0| \, \|\hat{g}_{1l}\|.$$

By the assumed boundedness of $\|\hat{g}_{1l}\|$, the last term is $O_p(|\tilde{\theta}_l - \theta_0|)$. The same decomposition holds for $\hat{\alpha}_{2l}(\tilde{\theta}_l) - \alpha_{02}(\theta_0)$. Summing the two bounds proves the displayed inequality for

$\|\hat{\alpha}_l(\tilde{\theta}_l) - \alpha_0(\theta_0)\|$, and combining it with

$$\|\hat{\gamma}_l - \gamma_0\| \leq \|\hat{g}_{1l} - g_{01}\| + \|\hat{g}_{2l} - g_{02}\|$$

gives the sufficient condition for Assumption 3.2(ii). □

Ridge-regularized sieves. This is the primitive example most closely aligned with the implementation used in Section 4. Let r_{0j} denote any one of the four conditional mean functions m_{01}, m_{02}, g_{01}, or g_{02}. Suppose r_{0j} admits an approximation $r_{K_j} = b'_{K_j}\beta_{K_j}$ in a K_j-dimensional sieve space with approximation error $a_{n,j} = \|r_{K_j} - r_{0j}\|$, the basis functions have uniformly bounded second moments, and the Gram matrix of b_{K_j} has eigenvalues bounded away from zero and infinity. Under these standard sieve regularity conditions, ridge-regularized least squares delivers

$$\|\hat{r}_{jl} - r_{0j}\| = O_p\left(a_{n,j} + \sqrt{K_j/n} + \lambda_{n,j}\|\beta_{K_j}\|_2\right)$$

for each regression; see Chen (2007). Proposition C.1 therefore implies that Assumptions 3.1 and 3.2 hold whenever these four regression rates vanish and their combined product with the generated-instrument rate is $o_p(n^{-1/2})$. In the balanced case with a root-n pilot and negligible ridge bias, the convenient sufficient condition is

$$a_{n,j} + \sqrt{K_j/n} = o_p(n^{-1/4}) \qquad \text{for each } r_{0j} \in \{m_{01}, m_{02}, g_{01}, g_{02}\}.$$

LASSO under approximate sparsity. Suppose each regression function $r_{0j} \in \{m_{01}, m_{02}, g_{01}, g_{02}\}$ admits an approximately sparse representation in a dictionary of size p_j, with effective sparsity s_j and approximation error $a_{n,j}$. Under standard restricted-eigenvalue conditions, LASSO or post-LASSO estimators satisfy

$$\|\hat{r}_{jl} - r_{0j}\| = O_p\left(a_{n,j} + \sqrt{s_j \log(p_j)/n}\right);$$

see Belloni et al. (2012) and the references collected in Chernozhukov, Newey, and Singh (2022a). Because the representers are explicit here, no separate sparsity condition for α_1 or α_2 is needed: Proposition C.1 converts these four conditional-mean rates directly into Assumptions 3.1 and 3.2. A convenient sufficient condition is that the pilot be root-n consistent and

$$a_{n,j} + \sqrt{s_j \log(p_j)/n} = o_p(n^{-1/4}) \qquad \text{for each } r_{0j} \in \{m_{01}, m_{02}, g_{01}, g_{02}\}.$$

Other flexible learners. The same logic applies more generally. Any estimator of the four conditional mean functions that delivers L^2 rates

$$r_{m1,n}, \quad r_{m2,n}, \quad r_{g1,n}, \quad r_{g2,n},$$

together with a pilot rate $r_{\theta,n} = |\tilde{\theta}_l - \theta_0|$, satisfies the high-level assumptions whenever

$$(r_{m1,n} + r_{m2,n} + r_{g1,n} + r_{g2,n} + r_{\theta,n})\,(r_{g1,n} + r_{g2,n}) = o(n^{-1/2}).$$

This covers neural networks, random forests, boosting, and other regularized nonparametric regressors under their respective standard smoothness, complexity, and stability conditions. The exact learner-specific primitive assumptions vary across methods, but the key point for the present estimator is that they need only guarantee these four conditional-mean rates. Section 3.2.3 of Chernozhukov, Newey, Quintas-Martinez, et al. (2024) surveys a broad class of such results.

Acknowledgments

I would like to thank Florian Gunsilius, Mel Stephens, Michal Kolesar, David Van Dijcke and Joel Robert Terschuur for their insightful comments and suggestions at various stages of this work. I would like to thank anonymous referees whose comments vastly improved the quality of this work. The usual disclaimer applies.

References

Ahrens, Achim et al. (2024). "ddml: Double/debiased Machine Learning in Stata". In: *The Stata Journal* 24.1, pp. 3–45. DOI: 10.1177/1536867X241233641.

Ai, Chunrong and Xiaohong Chen (2003). "Efficient estimation of models with conditional moment restrictions containing unknown functions". In: *Econometrica* 71.6, pp. 1795–1843.

Angrist, Joshua D and Brigham Frandsen (2022). "Machine labor". In: *Journal of Labor Economics* 40.S1, S97–S140.

Angrist, Joshua D, Guido W Imbens, et al. (1999). "Jackknife instrumental variables estimation". In: *Journal of Applied Econometrics* 14.1, pp. 57–67.

Bai, Jushan and Serena Ng (2010). "Instrumental variable estimation in a data rich environment". In: *Econometric Theory* 26.6, pp. 1577–1606.

Bald, Anthony et al. (2022). "Economics of foster care". In: *Journal of Economic Perspectives* 36.2, pp. 223–246.

Barua, Rashmi and Kevin Lang (2016). "Earnings of Apprentices: What Matters More – Workplace or Classroom Training?" In: *Journal of Human Capital* 10.4, pp. 470–500.

Belloni, Alexandre et al. (2012). "Sparse models and methods for optimal instruments with an application to eminent domain". In: *Econometrica* 80.6, pp. 2369–2429.

Bhuller, Manudeep et al. (2020). "Incarceration, Recidivism, and Employment". In: *Journal of Political Economy* 128.4, pp. 1269–1324. DOI: 10.1086/705330. eprint: https://doi.org/10.1086/705330. URL: https://doi.org/10.1086/705330.

Black, Bernard et al. (2018). "The effect of disability insurance receipt on mortality". In: *Memorandum 19/2012*. University College London.

Blandhol, Christine et al. (2025). "When is TSLS Actually LATE?" In: *Review of Economic Studies*. Forthcoming.

Bonhomme, Stéphane et al. (2026). "A Neyman-Orthogonalization Approach to the Incidental Parameter Problem". In: *arXiv preprint arXiv:2412.10304*.

Bruns-Smith, David (2025). "Two-Stage Machine Learning for Nonparametric Instrumental Variable Regression". In: *Working paper, Stanford University*.

Chao, John C et al. (2023). "Jackknife estimation of a cluster-sample IV regression model with many weak instruments". In: *Journal of Econometrics* 235.2, pp. 1747–1769. DOI: 10.1016/j.jeconom.2022.12.011. URL: https://doi.org/10.1016/j.jeconom.2022.12.011.

Chen, Jiafeng et al. (2021). "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models". In: *arXiv preprint arXiv:2011.06158*.

Chen, Xiaohong (2007). "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models". In: ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Handbook of Econometrics. Elsevier, pp. 5549–5632. DOI: https://doi.org/10.1016/S1573-4412(07)06076-X. URL: https://www.sciencedirect.com/science/article/pii/S157344120706076X.

Chernozhukov, Victor, Denis Chetverikov, et al. (2018). *Double/debiased machine learning for treatment and structural parameters*.

Chernozhukov, Victor, Juan Carlos Escanciano, et al. (2022). "Locally robust semiparametric estimation". In: *Econometrica* 90.4, pp. 1501–1535.

Chernozhukov, Victor, Christian Hansen, et al. (May 2015). "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments". In: *American Economic Review* 105.5, pp. 486–90. DOI: 10.1257/aer.p20151022. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20151022.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis (2020). "Adversarial estimation of riesz representers". In: *arXiv preprint arXiv:2101.00009*.

Chernozhukov, Victor, Whitney K Newey, Victor Quintas-Martinez, et al. (2021). "Automatic debiased machine learning via neural nets for generalized linear regression". In: *arXiv preprint arXiv:2104.14737*.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (2022a). "Automatic debiased machine learning of causal and structural effects". In: *Econometrica* 90.3, pp. 967–1027.

— (2022b). "Debiased machine learning of global and local parameters using regularized Riesz representers". In: *The Econometrics Journal* 25.3, pp. 576–601.

Chernozhukov, Victor, Whitney K. Newey, Victor Quintas-Martinez, et al. (2024). *Automatic Debiased Machine Learning via Riesz Regression*. arXiv: 2104.14737 [math.ST].

Chyn, Eric et al. (2025). "Examiner and Judge Designs in Economics: A Practitioner's Guide". In: *Journal of Economic Literature*. Forthcoming.

Coulibaly, Mohamed et al. (2024). "Sharp Testability of Monotonicity in Judge Designs". In: *arXiv preprint arXiv:2408.09098*.

Dobbie, Will and Jae Song (2015). "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection". In: *American economic review* 105.3, pp. 1272–1311.

Frandsen, Brigham, Lars Lefgren, et al. (2023). "Judging judge fixed effects". In: *American Economic Review* 113.1, pp. 253–277.

Frandsen, Brigham, Emily Leslie, et al. (2023). "Cluster Jackknife Instrumental Variables Estimation". URL: https://www.dropbox.com/scl/fi/po63fbmfgd65l60ihpbwt/Cluster_Jackknife20230807.pdf?rlkey=x0jfjw33am2pwp4w5c3eziubx&e=1&dl=0.

Geer, Sara van de et al. (2014). "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models". In: *The Annals of Statistics* 42.3, pp. 1166–1202.

Goldsmith-Pinkham, Paul et al. (2025). *Leniency Designs: An Operator's Manual*. Working Paper 34473. National Bureau of Economic Research.

Gross, Max and E Jason Baron (2022). "Temporary stays and persistent gains: The causal effects of foster care". In: *American Economic Journal: Applied Economics* 14.2, pp. 170–199.

Hahn, Jinyong (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica*, pp. 315–331.

Hahn, Jinyong and Jerry Hausman (2021). "Problems with the Control Variable Approach in Achieving Unbiased Estimates in Nonlinear Models in the Presence of Many Instruments". In: *Journal of Quantitative Economics* 19.Suppl. 1, S39–S58.

Hahn, Jinyong and Geert Ridder (2013). "Asymptotic Variance of Semiparametric Estimators With Generated Regressors". In: *Econometrica* 81.1, pp. 315–340. DOI: https://doi.org/10.3982/ECTA9609. eprint: https://onlinelibrary.wiley.com/doi/pdf/

10.3982/ECTA9609. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9609.

Hansen, Christian and Damian Kozbur (2014). "Instrumental variables estimation with many weak instruments using regularized JIVE". In: *Journal of Econometrics* 182.2, pp. 290–308. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2014.04.022. URL: https://www.sciencedirect.com/science/article/pii/S0304407614000918.

Hartford, Jason et al. (2017). "Deep IV: A Flexible Approach for Counterfactual Prediction". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 1414–1423.

Ichimura, Hidehiko and Whitney K Newey (2022). "The influence function of semiparametric estimators". In: *Quantitative Economics* 13.1, pp. 29–61.

Imbens, Guido W and Joshua D Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475.

Javanmard, Adel and Andrea Montanari (2014). "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression". In: *Journal of Machine Learning Research* 15.1, pp. 2869–2909.

Jochmans, Koen (2023). "Many (Weak) Judges in Judge-Leniency Designs". In.

Kamat, Vishal et al. (2023). "Identification in Multiple Treatment Models under Discrete Variation". In: *arXiv preprint arXiv:2307.06174*.

Kling, Jeffrey R. (June 2006). "Incarceration Length, Employment, and Earnings". In: *American Economic Review* 96.3, pp. 863–876. DOI: 10.1257/aer.96.3.863. URL: https://www.aeaweb.org/articles?id=10.1257/aer.96.3.863.

Kolesár, Michal (2013). "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity". In: *Unpublished working paper, Princeton University*.

Kolesár, Michal et al. (2025). "The Fragility of Sparsity". In: *arXiv preprint arXiv:2311.09299*.

Kosorok, Michael R (2008). *Introduction to empirical processes and semiparametric inference*. Vol. 61. Springer.

Mikusheva, Anna (2021). "Many Weak Instruments in Time Series Econometrics". In: *World Congress of Econometric Society. MIT*.

Mogstad, Magne and Alexander Torgovitsky (2024). "Instrumental Variables with Heterogeneous Treatment Effects". In: *Handbook of Labor Economics*. Forthcoming. Elsevier.

Mueller-Smith, Michael (2015). "The criminal and labor market impacts of incarceration". In: *Unpublished working paper* 18.

Newey, Whitney K (1994). "The asymptotic variance of semiparametric estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Newey, Whitney K. and Daniel McFadden (1994). "Chapter 36 Large sample estimation and hypothesis testing". In: vol. 4. Handbook of Econometrics. Elsevier, pp. 2111–2245. DOI: https://doi.org/10.1016/S1573-4412(05)80005-4. URL: https://www.sciencedirect.com/science/article/pii/S1573441205800054.

Perez-Izquierdo, Alvaro (2026). "Automatic Locally Robust GMM with Machine Learning-Generated Regressors". In: *arXiv preprint arXiv:2311.11127*.

Scheidegger, Cyrill et al. (2025). *Inference for Heterogeneous Treatment Effects with Efficient Instruments and Machine Learning*. DOI: 10.48550/arXiv.2503.03530. arXiv: 2503.03530 [stat.ME]. URL: https://doi.org/10.48550/arXiv.2503.03530.

Singh, Rahul and Liyang Sun (2024). "Double Robustness for Complier Parameters and a Semiparametric Test for Complier Characteristics". In: *The Econometrics Journal* 27.1, pp. 1–20.

Syrgkanis, Vasilis et al. (2019). *Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments*. arXiv: 1905.10176 [econ.EM].

Wiemann, Thomas (2023). "Optimal Categorical Instrumental Variables". In: *arXiv preprint arXiv:2311.17021*.

Zhang, Cun-Hui and Stephanie S Zhang (2014). "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 217–242.

Zivich, Paul N and Alexander Breskin (2021). "Machine Learning for Causal Inference: On the Use of Cross-fit Estimators". In: *Epidemiology* 32.3, pp. 393–401. DOI: 10.1097/EDE.0000000000001332.