The Econometrics Journal

Econometrics Journal (2024), volume **27**, pp. 1–20.
https://doi.org/10.1093/ectj/utad019

Double robustness for complier parameters and a semi-parametric test for complier characteristics

RAHUL SINGH[†] AND LIYANG SUN[‡]

[†]*MIT Economics, 50 Memorial Drive, Cambridge, MA 02142, USA.*
Email: r4hu1@mit.edu

[‡]*CEMFI Economics, 5 Calle Casado del Alisal, Madrid 28014, Spain.*
Email: lsun20@cemfi.es

First version received: 5 December 2022; final version accepted: 31 July 2023.

Summary: We propose a semi-parametric test to evaluate (a) whether different instruments induce subpopulations of compliers with the same observable characteristics, on average; and (b) whether compliers have observable characteristics that are the same as the full population, treated subpopulation, or untreated subpopulation, on average. The test is a flexible robustness check for the external validity of instruments. To justify the test, we characterise the doubly robust moment for Abadie's class of complier parameters, and we analyse a machine learning update to weighting that we call the automatic κ weight. We use the test to reinterpret Angrist and Evans' different local average treatment effect estimates obtained using different instrumental variables.

Keywords: *Instrumental variable*, *kappa weight*, *machine learning*, *semi-parametric efficiency*.

JEL codes: *C26*, *C14*, *C45*.

1. INTRODUCTION AND RELATED WORK

Average complier characteristics help to assess the external validity of any study that uses instrumental variable identification, such as Angrist and Evans (1998a), Angrist and Fernández-Val (2013a), Swanson and Hernán (2013), Baiocchi et al. (2014), and Marbach and Hangartner (2020). Whose treatment effects are we estimating when we use a particular instrument? We propose a semi-parametric hypothesis test, free of strong functional form restrictions, to evaluate (a) whether two different instruments induce subpopulations of compliers with the same observable characteristics, on average; and (b) whether compliers have observable characteristics that are the same as the full population, treated subpopulation, or untreated subpopulation, on average. It appears that no semi-parametric test previously exists for this important question about the external validity of instruments, despite the popularity of reporting average complier characteristics in empirical research, e.g., Abdulkadiroğlu et al. (2014, table 2). By developing this hypothesis test, we equip empirical researchers with a new robustness check.

Equipped with this new test, we replicate, extend, and test previous findings about the impact of childbearing on female labour supply. In a seminal paper, Angrist and Evans (1998a) use two different instrumental variables: twin births and same-sex siblings. The two instruments give rise to two substantially different local average treatment effect (LATE) estimates for the reduction in weeks worked due to a third child: −3.28 (0.63) and −6.36 (1.18), respectively, where the standard errors are in parentheses. Angrist and Fernández-Val (2013a) attribute the difference in LATE estimates to a difference in average complier characteristics, i.e., a difference in average covariates for instrument specific complier subpopulations, writing that 'twins compliers therefore are relatively more likely to have a young second-born and to be highly educated'. We find weak evidence in favour of the explanation that twins compliers are more likely to have a young second-born on average. We do not find evidence that twins compliers have a significantly different education level than same-sex compliers on average, but we do find a significant difference at the high end of the education distributions.

Our test is based on a new doubly robust estimator, which we call the automatic κ weight (Auto-κ). To prove the validity of the test, we characterise the doubly robust moment function for average complier characteristics, which appears to have been previously unknown. More generally, we study low dimensional complier parameters that are identified using a binary instrumental variable Z, which is valid conditional on a possibly high dimensional vector of covariates X. Angrist et al. (1996) prove that identification of LATE based on the instrumental variable does not require strong functional form restrictions; see Section 2. Using κ weighting, Abadie (2003) extends identification for a broad class of complier parameters. As our main theoretical result, we characterise the doubly robust moment function for this class of complier parameters by augmenting κ weighting with the classic Wald formula. Our main result answers the open question posed by Słoczyński and Wooldridge (2018) of how to characterise the doubly robust moment function for the full class, and it generalises the well-known result of Tan (2006), who characterises the doubly robust moment function for LATE. By characterising the doubly robust moment function for Abadie's (2003) class of complier parameters, we handle the new and economically important case of average complier characteristics.

The doubly robust moment function confers many favourable properties for estimation. As its name suggests, it provides double robustness to misspecification (Robins and Rotnitzky, 1995) as well as the mixed bias property (Chernozhukov et al., 2018; Rotnitzky et al., 2021). As such, it allows for estimation of models in which the treatment effect for different individuals may vary flexibly according to their covariates (Frölich, 2007; Ogburn et al., 2015). It also allows for nonlinear models (Abadie, 2003; Cheng et al., 2009), which are often appropriate when outcome Y and treatment D are binary, and therefore avoids the issue of negative weights in misspecified linear models (Blandhol et al., 2022). Moreover, it allows for model selection of covariates and their transformations using machine learning, as emphasised in the targeted machine learning (Van der Laan and Rubin, 2006; Zheng and Van der Laan, 2011; Luedtke and Van der Laan, 2016; Van der Laan and Rose, 2018) and debiased machine learning (Belloni et al., 2017; Chernozhukov et al., 2018, 2023; Chernozhukov et al., 2022) literature. A doubly robust estimator that combines both the κ weight and Wald formulations not only guards against misspecification, but also debiases machine learning. Finally, it is semi-parametrically efficient in many cases (Hasminskii and Ibragimov, 1979; Robinson, 1988; Bickel et al., 1993; Newey, 1994; Robins and Rotnitzky, 1995; Hong and Nekipelov, 2010).

This paper was previously circulated under a different title (Singh and Sun, 2019). Its structure is as follows. Section 2 defines the class of complier parameters from Abadie (2003). Section 3 summarises our main insight: the doubly robust moment for a complier parameter combines the familiar Wald and κ weight formulations. Section 4 formalises this insight for the full class of complier parameters. Section 5 develops the practical implication of our main insight: a semi-parametric test to evaluate differences in observable complier characteristics, which we use to revisit Angrist and Evans (1998a). Section 6 concludes. Appendix A proposes a machine learning estimator that we call the automatic κ weight (Auto-κ), which we use to implement our proposed test. Appendix B provides extensions: tests for the difference of (a) complier characteristic variances and (b) complier characteristic distributions over a finite support.

2. FRAMEWORK

Consider the effect of a binary treatment D on a scalar outcome Y in $\mathcal{Y}$, a subset of $\mathbb{R}$. Assume there is a binary instrumental variable Z available, as well as a potentially high dimensional covariate X in $\mathcal{X}$, a subset of $\mathbb{R}^{dim(X)}$. We observe n independent and identically distributed observations (W_i), $(i = 1, ..., n)$, where $W = (Y, D, Z, X^\top)^\top$ concatenates the random variables. Following the notation of Angrist et al. (1996), we denote by $Y^{(z,d)}$ the potential outcome under the intervention $Z = z$ and $D = d$. We denote by $D^{(z)}$ the potential treatment under the intervention $Z = z$. Compliers are the subpopulation for whom $D^{(1)} > D^{(0)}$. We place standard assumptions for identification.

ASSUMPTION 2.1 (INSTRUMENTAL VARIABLE IDENTIFICATION). *Assume the following conditions hold almost surely.*

(1) Independence: $\{Y^{(z,d)}\}, \{D^{(z)}\} \perp\!\!\!\perp Z \mid X$ *for* $d = 0, 1$ *and* $z = 0, 1$.
(2) Exclusion: $\mathrm{pr}\{Y^{(1,d)} = Y^{(0,d)} \mid X\} = 1$ *for* $d = 0, 1$.
(3) Overlap: $\pi_0(X) = \mathrm{pr}(Z = 1 \mid X)$ *is in (0,1).*
(4) Monotonicity: $\mathrm{pr}\{D^{(1)} \geq D^{(0)} \mid X\} = 1$ *and* $\mathrm{pr}\{D^{(1)} > D^{(0)} \mid X\} > 0$.

Independence states that the instrument Z is as good as randomly assigned conditional on covariates X. Exclusion imposes that the instrument Z only affects the outcome Y via the treatment D. We can therefore simplify notation: $Y^{(d)} = Y^{(1,d)} = Y^{(0,d)}$. Overlap ensures that there are no covariate values for which the instrument assignment is deterministic. Monotonicity rules out the possibility of defiers: individuals who will always pursue an opposite treatment status from their instrument assignment.

Angrist et al. (1996) prove identification of the local average treatment effect (LATE) using Assumption 2.1. Vytlacil (2002) shows that Assumption 2.1 implies the existence of a nonparametric latent index selection model that rationalises observed and counterfactual data. Abadie (2003) extends identification for a broad class of complier parameters.

DEFINITION 2.1 (GENERAL CLASS OF COMPLIER PARAMETERS; ABADIE, 2003). *Let* $g(y, d, x, \theta)$ *be a measurable, real valued function such that* $E\{g(Y, D, X, \theta)^2\} < \infty$ *for all* θ *in* Θ. *Consider complier parameters* θ_0 *implicitly defined by any of the following expressions:*

(1) $E\{g(Y^{(0)}, X, \theta) \mid D^{(1)} > D^{(0)}\} = 0$ *if and only if* $\theta = \theta_0$;
(2) $E\{g(Y^{(1)}, X, \theta) \mid D^{(1)} > D^{(0)}\} = 0$ *if and only if* $\theta = \theta_0$;
(3) $E\{g(Y, D, X, \theta) \mid D^{(1)} > D^{(0)}\} = 0$ *if and only if* $\theta = \theta_0$.

We refer to these expressions as the three possible cases for complier parameters.

For a given instrumental variable Z, one may define the average complier characteristics as a special case of Definition 2.1. This causal parameter summarises the observable characteristics of the complier subpopulation, who are induced to take up or refuse treatment D based on the instrument assignment Z. It is an important parameter to estimate because it aids the interpretation of LATE. As we will see in Section 5, this causal parameter can help to reconcile different LATE estimates obtained with different instruments.

DEFINITION 2.2 (AVERAGE COMPLIER CHARACTERISTICS). *Average complier characteristics are* $\theta_0 = E\{f(X) \mid D^{(1)} > D^{(0)}\}$ *for any measurable function* f *of covariate* X *that may have a finite dimensional, real vector value such that* $E\{f_j(X)^2\} < \infty$*, where* $f_j(X)$ *is the* j*th element of* $f(X)$*.*

3. KEY INSIGHT

3.1. Classic approaches: Wald formula and κ weight

We provide intuition for our key insight that a doubly robust moment for a complier parameter has two components: the Wald formula and the κ weight. For clarity, we focus on the familiar example of local average treatment effect (LATE) in this initial discussion: $\theta_0 = E\{Y^{(1)} - Y^{(0)} \mid D^{(1)} > D^{(0)}\}$. In subsequent sections, we study the entire class of complier parameters in Definition 2.1, including the new case of average complier characteristics.

Under Assumption 2.1, LATE can be identified as

$$\theta_0 = \frac{E\{E(Y \mid Z=1, X) - E(Y \mid Z=0, X)\}}{E\{E(D \mid Z=1, X) - E(D \mid Z=0, X)\}}$$

following Frölich (2007, thm. 1). We call this expression the expanded Wald formula.

The direct Wald approach involves estimating the reduced form regression $E(Y \mid Z, X)$ and first stage regression $E(D \mid Z, X)$, then plugging these estimates into the expanded Wald formula. Such an approach is called the plug-in, and it is valid only when both regressions are estimated with correctly specified and unregularised models. It is not a valid approach when either regression is incorrectly specified, leading to the name 'forbidden regression' (Angrist and Pischke, 2008). It is also invalid when the covariates are high dimensional and a regularised machine learning estimator is used to estimate either regression. The matching procedure of Frölich (2007) faces similar limitations.

In seminal work, Abadie (2003) proposes an alternative formulation in terms of the κ weights

$$\kappa^{(0)}(W) = (1-D)\frac{(1-Z) - \{1-\pi_0(X)\}}{\{1-\pi_0(X)\}\pi_0(X)}, \quad \kappa^{(1)}(W) = D\frac{Z - \pi_0(X)}{\{1-\pi_0(X)\}\pi_0(X)},$$

where $\pi_0(X) = \mathrm{pr}(Z = 1 \mid X)$ is the instrument propensity score. The κ weights have the property that

$$\theta_0 = \omega^{-1} E\{\kappa^{(1)}(W)Y - \kappa^{(0)}(W)Y\}, \quad \omega = E\left\{1 - \frac{D(1-Z)}{1-\pi_0(X)} - \frac{(1-D)Z}{\pi_0(X)}\right\}.$$

That is, the mean of the product of Y and $\kappa^{(d)}(W)$ gives, up to a scaling, the expected potential outcome $Y^{(d)}$ of compliers when treatment is $D = d$. Abadie (2003) also introduces a third weight $\kappa(W)$ for parameters that belong to the third case in Definition 2.1.

The κ weight approach would involve estimating the propensity score $\hat{\pi}$ and plugging this estimate into the κ weight formula. Intuitively, the κ weight approach is like a multistage inverse propensity weighting. Impressively, it remains agnostic about the functional form of the reduced form regression $E(Y \mid Z, X)$ and first stage regression $E(D \mid Z, X)$. It is valid only when $\hat{\pi}$ is estimated with a correctly specified and unregularised model. It is invalid if $\hat{\pi}$ is incorrectly specified or if covariates are high dimensional and a regularised machine learning estimator is used to estimate $\hat{\pi}$. Moreover, the inversion of $\hat{\pi}$ can lead to numerical instability in high dimensional settings.

3.2. Doubly robust moment for a special case

Next, we introduce the moment function and doubly robust moment function formulations of LATE. For the special case of LATE, these formulations were first derived by Tan (2006) with the goal of addressing misspecification of the regressions and the propensity score. Consider the expanded Wald formula. Rearranging and using the notation $V = (Y, D)^\top$ as a column vector, $\gamma_0(Z, X) = E(V \mid Z, X)$ as a vector valued regression, and $\big(1, -\theta\big)$ as a row vector, we arrive at the moment function formulation of LATE:

$$E\left[\big(1, -\theta\big)\{\gamma_0(1, X) - \gamma_0(0, X)\}\right] = 0 \text{ if and only if } \theta = \theta_0.$$

Denote the Horvitz–Thompson balancing weight as

$$\alpha_0(Z, X) = \frac{Z}{\pi_0(X)} - \frac{1-Z}{1-\pi_0(X)}, \quad \pi_0(X) = \mathrm{pr}(Z = 1 \mid X).$$

Tan (2006) shows that for LATE, the doubly robust moment function is

$$E\left[\big(1, -\theta\big)\{\gamma_0(1, X) - \gamma_0(0, X)\} + \alpha_0(Z, X)\big(1, -\theta\big)\{V - \gamma_0(Z, X)\}\right] = 0$$

if and only if $\theta = \theta_0$. The doubly robust formulation remains valid if either the vector valued regression γ_0 or propensity score π_0 is incorrectly specified.

3.3. A new synthesis that allows for machine learning

Our key observation is the connection between the κ weight and the balancing weight α_0. This simple observation will allow us to characterise the doubly robust moment function for a broad class of complier parameters, generalising Tan (2006) to the full class defined by Abadie (2003).

PROPOSITION 3.1 (κ WEIGHT AS BALANCING WEIGHT). *The κ weights can be rewritten as*

$$\kappa^{(0)}(W) = \alpha_0(Z, X)(D-1),\ \kappa^{(1)}(W) = \alpha_0(Z, X)D,\ \kappa(W) = 1 - \frac{D(1-Z)}{1-\pi_0(X)} - \frac{(1-D)Z}{\pi_0(X)}.$$

Proof: Observe that

$$\alpha_0(z, x) = \frac{z}{\pi_0(x)} - \frac{1-z}{1-\pi_0(x)} = \frac{z - \pi_0(x)}{\pi_0(x)\{1-\pi_0(x)\}}$$

which proves the expression for $\kappa^{(0)}$ and $\kappa^{(1)}$. Using these expressions, we have

$$\kappa(w) = \{1-\pi_0(x)\}\alpha_0(z, x)(d-1) + \pi_0(x)\alpha_0(z, x)d = 1 - \frac{d(1-z)}{1-\pi_0(x)} - \frac{(1-d)z}{\pi_0(x)}. \qquad \square$$

Next, we formalise the sense in which the balancing weight α_0 represents the functional $\gamma \mapsto E\left\{\big(1, -\theta\big)\gamma(1, X) - \gamma(0, X)\right\}$ that appears in the moment formulation of LATE and the extended Wald formula.

PROPOSITION 3.2 (BALANCING WEIGHT AS RIESZ REPRESENTER). *The balancing weight $\alpha_0(z, x)$ is the Riesz representer to the continuous linear functional $\gamma \mapsto E\{\gamma(1, X) - \gamma(0, X)\}$,* *i.e., for all γ such that $E\{\gamma(Z, X)\}^2 < \infty$,*

$$E\{\gamma(1, X) - \gamma(0, X)\} = E\{\alpha_0(Z, X)\gamma(Z, X)\}.$$

Similarly, $Z/\pi_0(X)$ is the Riesz representer to the continuous linear functional $\gamma \mapsto E\{\gamma(1, X)\}$, and $(1 - Z)/\{1 - \pi_0(X)\}$ is the Riesz representer to the continuous linear functional $\gamma \mapsto E\{\gamma(0, X)\}$.

Proof: This result is well known in semi-parametrics and it follows from the law of iterated expectations. See, e.g., Hernán and Robins (2020). □

An immediate consequence of Proposition 3.2 is that

$$E\left\{\big(1, -\theta\big)\gamma(1, X) - \gamma(0, X)\right\} = E\left\{\alpha_0(Z, X)\big(1, -\theta\big)\gamma(Z, X)\right\} \text{ for any } \gamma.$$

In summary, Proposition 3.1 shows that the κ weight is a re-parametrisation of the balancing weight α_0. Meanwhile, Proposition 3.2 shows that the balancing weight appears in the Riesz representer to the moment formulation of LATE, i.e., the expanded Wald formula. We conclude that the κ weight is essentially the Riesz representer to the Wald formula. In seminal work, Newey (1994) demonstrates that a doubly robust moment is constructed from a moment formulation and its Riesz representer. Therefore the doubly robust moment for complier parameters must combine the Wald formula and the κ weight.

With the general doubly robust moment function, one can propose flexible, semi-parametric tests for complier parameters. In particular, the semi-parametric tests may involve regularised machine learning for flexible estimation and model selection of (a) the regression $\hat{\gamma}$ in a way that approximates nonlinearity and heterogeneity, and (b) the balancing weight $\hat{\alpha}$ in a way that guarantees balance. In Section 5, we instantiate such a test to compare observable characteristics of compliers.

As explained in Appendix A, we avoid the numerically unstable step of estimating and inverting $\hat{\pi}$ that appears in Tan (2006), Belloni et al. (2017), and Chernozhukov et al. (2018). We replace it with the numerically stable step of estimating $\hat{\alpha}$ directly, extending techniques of Chernozhukov et al. (2022a) to the instrumental variable setting. We call this extension automatic κ weighting (Auto-κ), and demonstrate how it applies to the new and economically important case of average complier characteristics. In Appendix B, we extend our framework to additional new cases: complier characteristic variances and complier characteristic distributions over a finite support.

In summary, our main theoretical result allows us to combine the classic Wald and κ weight formulations for the entire class of complier parameters in Definition 2.1, including average complier characteristics, while also updating them to use machine learning.

4. THE DOUBLY ROBUST MOMENT

We now state our main theoretical result, which is the doubly robust moment for the class of complier parameters in Definition 2.1. This result formalises the intuition of Section 3, and it justifies the hypothesis test in Section 5. It is convenient to divide the main result into two statements for clarity. Theorem 4.1 handles the first and second cases in Definition 2.1, while Theorem 4.2 handles the third case in Definition 2.1.

THEOREM 4.1 (CASES 1 AND 2). *Suppose Assumption 2.1 holds. Let $g(y, d, x, \theta)$ be a measurable, real valued function such that $E\{g(Y, D, X, \theta)^2\} < \infty$ for all θ in Θ.*

(1) If θ_0 is defined by $E[g\{Y^{(0)}, X, \theta_0\} \mid D^{(1)} > D^{(0)}] = 0$, let $v(w, \theta) = (d-1)g(y, x, \theta)$.
(2) If θ_0 is defined by $E[g\{Y^{(1)}, X, \theta_0\} \mid D^{(1)} > D^{(0)}] = 0$, let $v(w, \theta) = dg(y, x, \theta)$.

Then the doubly robust moment function ψ for θ_0 is of the form

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta), \quad m(w, \gamma, \theta) = \gamma(1, x, \theta) - \gamma(0, x, \theta),$$
$$\phi(w, \gamma, \alpha, \theta) = \alpha(z, x)\{v(w, \theta) - \gamma(z, x, \theta)\},$$

where $\gamma_0(z, x, \theta) = E\{v(W, \theta) \mid z, x\}$ is a vector valued regression and $\alpha_0(z, x) = z/\pi_0(x) - (1-z)/\{1-\pi_0(x)\}$ is the Riesz representer of the functional $\gamma \mapsto E\{\gamma(1, X, \theta) - \gamma(0, X, \theta)\}$.

Proof: Consider the first case. Under Assumption 2.1, we can appeal to Abadie (2003, thm. 3.1):

$$0 = E[g\{Y^{(0)}, X, \theta_0\} \mid D^{(1)} > D^{(0)}] = \frac{E\{\kappa^{(0)}(W)g(Y, X, \theta_0)\}}{\text{pr}\{D^{(1)} > D^{(0)}\}}.$$

Hence

$$0 = E\{\kappa^{(0)}(W)g(Y, X, \theta_0)\} = E\{\alpha_0(Z, X)(D-1)g(Y, X, \theta_0)\} = E\{\alpha_0(Z, X)v(W, \theta_0)\}$$
$$= E\{\alpha_0(Z, X)\gamma_0(Z, X, \theta_0)\} = E\{\gamma_0(1, X, \theta_0) - \gamma_0(0, X, \theta_0)\}$$

appealing to the previous statement, Proposition 3.1, the definition of $v(W, \theta_0)$, the law of iterated expectations, and Proposition 3.2. Likewise for the second case. □

In the doubly robust moment function $\psi(w, \gamma, \alpha, \theta) = m(w, \gamma, \theta) + \phi(w, \gamma, \alpha, \theta)$, we generalise our insight from Section 3. The first term $m(w, \gamma, \theta)$ is essentially a generalised Wald formula. The second term $\phi(w, \gamma, \alpha, \theta)$ is essentially a product between the κ weight and a generalised regression residual. In the language of semi-parametrics, we *augment* the κ weight with the Wald formula. Equivalently, we *debias* the Wald formula with the κ weight.

The doubly robust moment function ψ remains valid if either γ_0 or α_0 is misspecified:

$$0 = E\{\psi(W, \gamma, \alpha_0, \theta_0)\} = E\{\psi(W, \gamma_0, \alpha, \theta_0)\} \text{ for any } \gamma, \alpha.$$

In the former expression, γ_0 may be misspecified yet ψ remains valid as an estimating equation. In the latter, α_0 may be misspecified yet ψ remains valid as an estimating equation. Theorem 4.1 demonstrates that all complier parameters in cases 1 and 2 of Definition 2.1 have a doubly robust moment function ψ with a common structure. As such, we are able to analyse all of these causal parameters with the same argument. Case 3 of Definition 2.1 is more involved, but we show that it shares this common structure.

THEOREM 4.2 (CASE 3). *Suppose Assumption 2.1 holds. Let $g(y, d, x, \theta)$ be a measurable, real valued function such that $E\{g(Y, D, X, \theta)^2\} < \infty$ for all θ in Θ. If θ_0 is defined by the moment condition $E\{g(Y, D, X, \theta_0) \mid D^{(1)} > D^{(0)}\} = 0$, then the doubly robust moment function for θ_0 is of the form*

$$\psi(w, \tilde{\gamma}, \tilde{\alpha}, \theta) = m(w, \tilde{\gamma}, \theta) + \phi(w, \tilde{\gamma}, \tilde{\alpha}, \theta), \; m(w, \tilde{\gamma}, \theta) = \gamma(z, x, \theta) - \gamma^0(1, x, \theta) - \gamma^1(0, x, \theta)$$
$$\phi(w, \tilde{\gamma}, \tilde{\alpha}, \theta) = \{g(y, d, x, \theta) - \gamma(z, x, \theta)\} - \alpha^0(z, x)\{(1-d)g(y, d, x, \theta) - \gamma^0(z, x, \theta)\}$$
$$- \alpha^1(z, x)\{dg(y, d, x, \theta) - \gamma^1(z, x, \theta)\},$$

where $\tilde{\gamma}$ *concatenates* $(\gamma, \gamma^0, \gamma^1)$ *and* $\tilde{\alpha}$ *concatenates* (α^0, α^1)*. These functions are*

$$\gamma_0(z, x, \theta) = E\{g(Y, D, X, \theta) \mid z, x\}, \quad \gamma_0^0(z, x, \theta) = E\{(1 - D)g(Y, D, X, \theta) \mid z, x\},$$
$$\gamma_0^1(z, x, \theta) = E\{Dg(Y, D, X, \theta) \mid z, x\}, \;\; \alpha_0^0(z, x) = z/\pi_0(x), \;\; \alpha_0^1(z, x) = (1 - z)/\{1 - \pi_0(x)\}.$$

Proof: The argument is similar to the proof of Theorem 4.1. Under Assumption 2.1, we can appeal to Abadie (2003, thm. 3.1):

$$0 = E\{g(Y, D, X, \theta_0) \mid D^{(1)} > D^{(0)}\} = \frac{E\{\kappa(W)g(Y, D, X, \theta_0)\}}{\mathrm{pr}\{D^{(1)} > D^{(0)}\}}.$$

Hence

$$\begin{aligned}
0 &= E\{\kappa(W)g(Y, D, X, \theta_0)\} \\
&= E\left\{g(Y, D, X, \theta_0) - \frac{Z}{\pi_0(X)}(1 - D)g(Y, D, X, \theta_0) - \frac{1 - Z}{1 - \pi_0(X)}Dg(Y, D, X, \theta_0)\right\} \\
&= E\left\{\gamma_0(Z, X, \theta_0) - \frac{Z}{\pi_0(X)}\gamma_0^0(Z, X, \theta_0) - \frac{1 - Z}{1 - \pi_0(X)}\gamma_0^1(Z, X, \theta_0)\right\} \\
&= E\{\gamma_0(Z, X, \theta_0) - \gamma_0^0(1, X, \theta_0) - \gamma_0^1(0, X, \theta_0)\}
\end{aligned}$$

appealing to the previous statement, Proposition 3.1, the definitions of $(\gamma_0, \gamma_0^0, \gamma_0^1)$ together with the law of iterated expectations, and Proposition 3.2. □

This time, the doubly robust moment function ψ remains valid if either $\tilde{\gamma}_0$ or $\tilde{\alpha}_0$ is misspecified, i.e.,

$$0 = E\{\psi(W, \tilde{\gamma}, \tilde{\alpha}_0, \theta_0)\} = E\{\psi(W, \tilde{\gamma}_0, \tilde{\alpha}, \theta_0)\} \text{ for any } \tilde{\gamma}, \tilde{\alpha}.$$

In the former expression, $\tilde{\gamma}_0$ may be misspecified yet ψ remains valid as an estimating equation. In the latter, $\tilde{\alpha}_0$ may be misspecified yet ψ remains valid as an estimating equation.

In Section 5, we translate this general characterisation of the doubly robust moment into a practical hypothesis test to evaluate the external validity of instruments. In Appendix A, we translate this general characterisation into general machine learning estimators for complier parameters, which we use to implement the hypothesis test. In particular, we consider direct estimation of the balancing weight, a procedure that we call automatic κ weighting (Auto-κ). In Appendix B, we translate this general characterisation into further hypothesis tests.

5. A HYPOTHESIS TEST TO COMPARE OBSERVABLE CHARACTERISTICS

5.1. Corollaries for average complier characteristics

As a corollary, we characterise the doubly robust moment for average complier characteristics, which appears to have been previously unknown. Using the new doubly robust moment, we propose a hypothesis test, free of strong functional form restrictions, to evaluate (a) whether two different instruments induce subpopulations of compliers with the same observable characteristics, on average; and (b) whether compliers have observable characteristics that are the same as the full population, treated subpopulation, or untreated subpopulation, on average.

COROLLARY 5.1 (AVERAGE COMPLIER CHARACTERISTICS). *The doubly robust moment for average complier characteristics is*

$$\psi(w, \gamma, \alpha, \theta) = A(\theta)\{\gamma(1, x) - \gamma(0, x)\} + \alpha(z, x)A(\theta)\{v - \gamma(z, x)\}, \quad A(\theta) = \big(I, -\theta\big),$$

where $v = \{df(x)^\top, d\}^\top$, $\gamma_0(z, x) = E(V \mid z, x)$, *and* $\alpha_0(z, x) = z/\pi_0(x) - (1 - z)/\{1 - \pi_0(x)\}$.

Proof: The result is a special case of Corollary A.1 in Appendix A. □

Suppose we wish to test the null hypothesis that two different instruments Z_1 and Z_2 induce complier subpopulations with the same observable characteristics on average. Denote by $\hat{\theta}_1$ and $\hat{\theta}_2$ the estimators for average complier characteristics using the different instruments Z_1 and Z_2, respectively. One may construct machine learning estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ based on the doubly robust moment function in Corollary 5.1. In Appendix A, we instantiate automatic κ weight (Auto-κ) estimators of this type. The following procedure allows us to test the null hypothesis from some estimator $\hat{C}$ for the asymptotic variance C of $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$. Appendix A provides an explicit variance estimator $\hat{C}$ based on Auto-κ.

ALGORITHM 5.1 (TEST FOR DIFFERENCE OF AVERAGE COMPLIER CHARACTERISTICS). *Given $\hat{\theta}$ and $\hat{C}$, which may be based on Auto-κ as in Appendix A,*

STEP *1. Calculate the statistic* $T = n(\hat{\theta}_1 - \hat{\theta}_2)^\top (R\hat{C}R^\top)^{-1}(\hat{\theta}_1 - \hat{\theta}_2)$ *where* $R = \big(I, -I\big)$.
STEP *2. Compute the value c_a as the $(1 - a)$ quantile of $\chi^2\{dim(\theta_1)\}$.*
STEP *3. Reject the null hypothesis if $T > c_a$.*

Algorithm 5.1 can also test the null hypothesis that compliers have observable characteristics that are the same as the full population on average. $\hat{\theta}_1$ is as before, $\hat{\theta}_2 = n^{-1}\sum_{i=1}^n f(X_i)$, and $\hat{C}$ updates accordingly. The same is true for comparisons with the treated and untreated subpopulations. For example, for the former, $\hat{\theta}_1$ is as before, $\hat{\theta}_2 = (\sum_{i=1}^n D_i)^{-1}\sum_{i=1}^n D_i f(X_i)$, and $\hat{C}$ updates accordingly. In summary, the test assesses how similar, in terms of observable characteristics, the complier subpopulation is to other (sub)populations of interest: the complier subpopulation for a different instrument, the full population, the treated subpopulation, or the untreated subpopulation. The null hypothesis has the dimension $dim(\theta_1)$, which is finite following Definition 2.2. Future work may consider high dimensional complier characteristics.

The test sheds some light on the robustness and external validity of policy evaluation when using instruments. If the complier subpopulations in various studies are dissimilar to each other, then the policy conclusions of those studies may not be robust: different choices of instruments may lead to divergent policy conclusions. If the complier subpopulation in a study is dissimilar to the population of policy interest, then that study may lack external validity: its policy conclusions may not hold for the relevant population.

The test focuses on only observable characteristics, so it is a partial answer to the question of whose treatment effects are being estimated when using a particular instrument. It complements estimation of the fraction of compliers in the sample; the fraction of compliers could be small, yet the compliers could have observable characteristics that are similar to the (sub)population of interest.

COROLLARY 5.2 (TEST FOR DIFFERENCE OF AVERAGE COMPLIER CHARACTERISTICS). *If* $n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, C)$ *and* $\hat{C} = C + o_p(1)$, *then the hypothesis test in Algorithm 5.1 falsely rejects the null hypothesis H_0 with probability approaching the nominal level, i.e.,* $\mathrm{pr}(T > c_a \mid H_0) \to a$.

Proof: The result is immediate from Newey and McFadden (1994, sect. 9). □

Corollary 5.2 is our main practical result: justification of a flexible hypothesis test to evaluate a difference in average complier characteristics. It appears that no semi-parametric test previously exists for this important question about the external validity of instruments. By developing this hypothesis test, we equip empirical researchers with a new robustness check. This practical result follows as a consequence of our main insight in Section 3 and our main theoretical result in Section 4. In Appendix A, we verify the conditions of Corollary 5.2 for Auto-κ under additional weak regularity assumptions.

In terms of power, the test based on Auto-κ estimation is asymptotically efficient (Van der Vaart, 2000) because the Auto-κ estimator for average complier characteristics is semi-parametrically efficient. See Appendix A for formal justification. Specifically, in Appendix A, we verify that average complier characteristics belong to a subclass of complier parameters with affine moments. For this subclass, the doubly robust moment coincides with the semi-parametrically efficient score (Hahn, 1998).

Our framework naturally extends to test (a) a finite number of moments of complier characteristics and (b) a finitely supported distribution of complier characteristics. For uncentred moments, the extension is immediate: simply take the function f in Definition 2.2 to be the polynomial corresponding to the desired moments. For centred moments, the extension is straightforward, which we demonstrate in Appendix B by presenting a test for the difference of complier characteristic variances. We also present a test for the difference of complier characteristic distributions, over a finite support.

5.2. *Empirical application*

With this practical result, we revisit a classic empirical paper in labour economics to test whether two different instruments induce different average complier characteristics. Angrist and Evans (1998a) estimate the impact of childbearing D on female labour supply Y in a sample of 394,840 mothers, aged 21–35 with at least two children, from the 1980 census (Angrist and Evans 1998b; Angrist and Fernández-Val 2013b). The first instrument Z_1 is twin births: Z_1 indicates whether the mother's second and third children were twins. The second instrument Z_2 is same-sex siblings: Z_2 indicates whether the mother's initial two children were siblings with the same sex. The authors reason that both (Z_1, Z_2) are quasi-random events that induce having a third child.

The two instruments give rise to two LATE estimates for the reduction in weeks worked due to a third child: -3.28 (0.63) for Z_1 and -6.36 (1.18) for Z_2, where the standard errors are in parentheses. Angrist and Fernández-Val (2013a) attribute the difference in LATE estimates to a difference in average complier characteristics, i.e., a difference in average covariates for instrument specific complier subpopulations. The authors use parametric κ weights, report point estimates without standard errors, and conclude that 'twins compliers therefore are relatively more likely to have a young second-born and to be highly educated'.

We replicate, extend, and test these previous findings. Using parametric κ weights, Angrist and Fernández-Val (2013a) estimate $\pi_0(X)$ using a logistic model with polynomials of continuous covariates. In our semi-parametric Auto-κ approach, we expand the dictionary to higher-order polynomials, include interactions between the instrument and covariates, and directly estimate and regularise the balancing weights. Crucially, our main result allows us to conduct inference, and to test whether the instruments Z_1 and Z_2 induce differences in the observable complier characteristics suggested by previous work.

Table 1. Comparison of average complier characteristics.

	Average age of second child				Average schooling of mother			
	Twins	Same	2 sided	1 sided	Twins	Same	2 sided	1 sided
κ weight	5.51	7.14	–	–	12.43	12.07	–	–
Auto-κ	4.58	7.00	0.14	0.07	9.78	12.10	0.53	0.27
(SE)	(0.72)	(1.46)	–	–	(2.44)	(2.78)	–	–

Notes: SE, standard error; Auto-κ, automatic κ weighting. See Appendix A and Online Appendix S5 for estimation details.

Table 1 summarises results. In Columns 1, 2, 5, and 6, we find similar point estimates to Angrist and Fernández-Val (2013a), given in Row 1. Columns 3, 4, 7, and 8 report p-values for tests of the null hypothesis that average complier characteristics are equal for the twins and same-sex instruments. We find weak evidence in favour of the explanation that twins compliers are more likely to have a young second-born. We do not find evidence that twins compliers have a significantly different education level than same-sex compliers in terms of the average years of mother's schooling. In Appendix B, we discretise mother's schooling into a categorical variable, then test for a difference in distributions of education categories. We find evidence that twins compliers are more likely to be college graduates, corroborating the conclusions of Angrist and Fernández-Val (2013a).

6. CONCLUSION

We propose a semi-parametric test to evaluate (a) whether two different instruments induce subpopulations of compliers with the same observable characteristics, on average; and (b) whether compliers have observable characteristics that are the same as the full population, treated subpopulation, or untreated subpopulation, on average. This hypothesis test is flexible and practical, shedding light on the difference in LATE estimates that Angrist and Evans (1998a) obtain when using two different instruments. As a contribution to semi-parametric theory, we characterise the doubly robust moment function for the entire class of complier parameters from Abadie (2003), answering an open question in order to handle the new and economically important case of average complier characteristics. As a contribution to applied econometrics, we propose and analyse a machine learning update to κ weighting that we call the automatic κ weight (Auto-κ).

ACKNOWLEDGEMENTS

We thank Alberto Abadie, Anish Agarwal, Isaiah Andrews, Joshua Angrist, Abhijit Banerjee, Victor Chernozhukov, Peng Ding, Avi Feller, Brigham Frandsen, Guido Imbens, Patrick Kline, Lihua Lei, Anna Mikusheva, Whitney Newey, Garima Sharma, Tymon Słoczyński, and Suhas Vijaykumar for helpful comments. The paper also benefited from the comments of conference and seminar audiences at NeurIPS CausalML Workshop 2019, Econometric Society World Congress 2020, NABE Tech Economics Conference 2020, California Econometrics Conference 2022, LMU Munich, UCL, and African Econometrics Society Meeting 2023. Both authors thank the Jerry Hausman Dissertation Fellowship. Part of this work was done while Rahul Singh was visiting the

Simons Institute for the Theory of Computing. Liyang Sun gratefully acknowledges support from the Institute of Education Sciences, US Department of Education, through Grant R305D200010, and Ayudas Juan de la Cierva Formación.

REFERENCES

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–63.

Abdulkadiroğlu, A., J. Angrist and P. Pathak (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica 82*(1), 137–96.

Angrist, J. D. and W. N. Evans (1998a). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–77.

Angrist, J. D. and W. N. Evans (1998b). Children and their parents' labor supply: Evidence from exogenous variation in family size: Replication package. *American Economic Review 88*(3), Data deposited at https://economics.mit.edu/people/faculty/josh--angrist/angrist--data--archive.

Angrist, J. D. and I. Fernández-Val (2013a). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In D. Acemoglu, M. Arellano and E. Dekel (Eds.), *Advances in Economics and Econometrics*, 401–34. Cambridge: Cambridge University Press.

Angrist, J. D. and I. Fernández-Val (2013b). ExtrapoLATE-ing: External validity and overidentification in the LATE framework: Replication package. In *Advances in Economics and Econometrics*, Data deposited at https://sites.bu.edu/ivanf/research/.

Angrist, J. D., G. W. Imbens and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–55.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Athey, S., G. W. Imbens and S. Wager (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Baiocchi, M., J. Cheng and D. S. Small (2014). Instrumental variable methods for causal inference. *Statistics in Medicine 33*(13), 2297–340.

Belloni, A., V. Chernozhukov, I. Fernández-Val and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–98.

Belloni, A., V. Chernozhukov and L. Wang (2014). Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics 42*(2), 757–88.

Bickel, P. J., C. A. Klaassen, Y. Ritov and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, vol. 4. Baltimore, MD: Johns Hopkins University Press.

Blandhol, C., J. Bonney, M. Mogstad and A. Torgovitsky (2022). When is TSLS actually LATE? Technical report, National Bureau of Economic Research.

Cheng, J., D. S. Small, Z. Tan and T. R. Ten Have (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika 96*(1), 19–36.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey and J. M. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*(1), C1–68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–35.

Chernozhukov, V., W. K. Newey and R. Singh (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Chernozhukov, V., W. K. Newey and R. Singh (2022b). De-biased machine learning of global and local parameters using regularized Riesz representers. *Econometrics Journal 25*(3), 576–601.

Chernozhukov, V., W. K. Newey and R. Singh (2023). A simple and general debiased machine learning theorem with finite sample guarantees. *Biometrika 110*(1), 257–64.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*, 315–31.

Hasminskii, R. Z. and I. A. Ibragimov (1979). On the nonparametric estimation of functionals. In P. Mandl and M. Hušková (Eds.), *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 41–51. Amsterdam: North-Holland.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton, FL: CRC.

Hirshberg, D. A. and S. Wager (2021). Augmented minimax linear estimation. *Annals of Statistics 49*(6), 3206–27.

Hong, H. and D. Nekipelov (2010). Semiparametric efficiency in nonlinear LATE models. *Quantitative Economics 1*(2), 279–304.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics 15*, 1548–62.

Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics 44*(2), 713.

Marbach, M. and D. Hangartner (2020). Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis 28*(3), 435–44.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*, 1349–82.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engl and D. McFadden (Eds.), *Handbook of Econometrics*, vol. 4, 2111–245. Amsterdam: Elsevier.

Ogburn, E. L., A. Rotnitzky and J. M. Robins (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(2), 373–96.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–9.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*, 931–54.

Rotnitzky, A., E. Smucler and J. M. Robins (2021). Characterization of parameters with a mixed bias property. *Biometrika 108*(1), 231–8.

Singh, R. and L. Sun (2019). De-biased machine learning in instrumental variable models for treatment effects. *arXiv: Statistics, Machine Learning* 1909.05244.

Słoczyński, T. and J. M. Wooldridge (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory 34*(1), 112–33.

Swanson, S. A. and M. A. Hernán (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology 24*(3), 370–4.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association 101*(476), 1607–18.

Tsybakov, A. B. (2012). *Introduction to Nonparametric Estimation*. New York: Springer.

Van der Laan, M. J. and S. Rose (2018). *Targeted Learning in Data Science*. New York: Springer.

Van der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics 2*(1), 11.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, vol. 3. Cambridge: Cambridge University Press.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–41.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In M. van der Laan and S. Rose (Eds.), *Targeted Learning*, 459–74. New York: Springer.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–22.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Petra Todd handled this manuscript.

APPENDIX A: AUTOMATIC $\boldsymbol{\kappa}$ WEIGHTS

A.1. Estimation

In Section 4, we present our main theoretical result: the doubly robust moment function for the class of complier parameters in Definition 2.1. In this section, we propose a machine learning estimator based on this doubly robust moment function, which we call automatic κ weighting (Auto-κ). We verify the conditions of Corollary 5.2 using Auto-κ. In doing so, we provide a concrete end-to-end procedure to test whether two different instruments induce subpopulations of compliers with the same observable characteristics.

Debiased machine learning (Chernozhukov et al., 2018, 2023; Chernozhukov et al., 2022) is a meta estimation procedure that combines doubly robust moment functions (Robins and Rotnitzky, 1995) with sample splitting (Klaassen, 1987). Given the doubly robust moment function of some causal parameter of interest as well as machine learning estimators $(\hat{\gamma}, \hat{\alpha})$ for its nonparametric components, debiased machine learning generates an estimator of the causal parameter.

ALGORITHM A.1 (DEBIASED MACHINE LEARNING). *Partition the sample into subsets (I_ℓ), $(\ell = 1, ..., L)$.*

STEP *1. For each ℓ, estimate $\hat{\gamma}_{-\ell}$ and $\hat{\alpha}_{-\ell}$ from observations not in I_ℓ.*
STEP *2. Estimate $\hat{\theta}$ as the solution to $n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta)|_{\theta=\hat{\theta}} = 0$.*

In Theorems 4.1 and 4.2, we characterise the doubly robust moment function ψ for complier parameters. What remains is an account of how to estimate the vector valued regression $\hat{\gamma}$ and the balancing weight $\hat{\alpha}$. Our theoretical results are agnostic about the choice of $(\hat{\gamma}, \hat{\alpha})$ as long as they satisfy the rate conditions in Assumption A.1. For example, $\hat{\gamma}$ could be a neural network.

For the balancing weight estimator $\hat{\alpha}$, we adapt the regularised Riesz representer of Chernozhukov et al. (2022a), though one could similarly adapt the minimax balancing weight of Hirshberg and Wager (2021). This aspect of the procedure departs from the explicit inversion of the propensity score in Tan (2006); Belloni et al. (2017); Chernozhukov et al. (2018), and it improves numerical stability, which we demonstrate though comparative simulations in Online Appendix S4. In particular, we project the balancing weight $\alpha_0(Z, X)$ onto the p dimensional dictionary of basis functions $b(Z, X)$. A high dimensional dictionary allows for flexible approximation, which we discipline with ℓ_1 regularisation.

ALGORITHM A.2 (REGULARIZED BALANCING WEIGHT). *Let $I_{-\ell}$ be the complement of I_ℓ, and let $n_\ell = |I_\ell|$. Based on the observations in $I_{-\ell}$,*

STEP *1. Calculate $p \times p$ matrix $\hat{G}_{-\ell} = (n - n_\ell)^{-1} \sum_{i \in I_{-\ell}} b(Z_i, X_i) b(Z_i, X_i)^\top$.*
STEP *2. Calculate $p \times 1$ vector $\hat{M}_{-\ell} = (n - n_\ell)^{-1} \sum_{i \in I_{-\ell}} b(1, X_i) - b(0, X_i)$.*
STEP *3. Set $\hat{\alpha}_{-\ell}(Z, X) = b(Z, X)^\top \hat{\rho}_{-\ell}$ where $\hat{\rho}_{-\ell} = \operatorname{argmin}_\rho \rho^\top \hat{G}_{-\ell} \rho - 2\rho^\top \hat{M}_{-\ell} + 2\lambda_n |\rho|_1$.*

The regularisation parameter λ_n is determined by an iterative tuning procedure described in Online Appendix S3.

As summarised by Chernozhukov et al. (2022b), regularised linear combinations of conventional basis functions $b(Z, X)$ may be used to approximate certain function classes well. For example, Tsybakov (2012) shows that Fourier bases approximate Sobolev balls, and Belloni et al. (2014) show that Fourier bases approximate rearranged Sobolev balls.

We refer to our proposed estimator, which combines the doubly robust moment function from Theorems 4.1 and 4.2 with the meta procedure in Algorithm A.1 and the regularised balancing weights in Algorithm A.2, as automatic κ weighting (Auto-κ) for complier parameters. The new doubly robust moment in Corollary 5.1 means that Auto-κ applies to the new and economically important case of average complier characteristics.

A.2. Approximate balance

Auto-κ confers a finite sample guarantee of balance on average. Consider the dictionary of basis functions $b(z, x)^\top = \{zq(x)^\top, (1-z)q(x)^\top\}$ and the corresponding partition of the coefficient $\rho^\top = [\{\rho^{(z=1)}\}^\top, \{\rho^{(z=0)}\}^\top]$.

PROPOSITION A.1 (APPROXIMATE BALANCE). *Auto-κ with regularisation λ_n yields*

$$\left| \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} q(X_i) - \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} q(X_i) Z_i \cdot \hat{\omega}^{(z=1)}_{-\ell,i} \right|_\infty \leq \lambda_n,$$

$$\left| \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} q(X_i) - \frac{1}{n-n_\ell} \sum_{i \in I_{-\ell}} q(X_i)(1-Z_i) \cdot \hat{\omega}^{(z=0)}_{-\ell,i} \right|_\infty \leq \lambda_n,$$

for all n, where $\hat{\omega}^{(z=1)}_{-\ell,i} = q(X_i)^\top \hat{\rho}^{(z=1)}_{-\ell}$ and $\hat{\omega}^{(z=0)}_{-\ell,i} = q(X_i)^\top \hat{\rho}^{(z=0)}_{-\ell}$.

Proof: The first order condition gives $|\hat{M}_{-\ell} - \hat{G}_{-\ell}\hat{\rho}_{-\ell}|_\infty \leq \lambda_n$. □

Proposition A.1 shows that the weights $\{\hat{\omega}^{(z=1)}_{\ell,i}, \hat{\omega}^{(z=0)}_{\ell,i}\}$ serve to approximately balance the overall sample average with the sample average of the group that is assigned the instrument ($Z = 1$) and the sample average of the group that is not assigned the instrument ($Z = 0$), across each basis function of the dictionary q. The result is similar to the balancing conditions of Zubizarreta (2015) and Athey et al. (2018). Auto-κ automatically calculates these weights. This property does not hold for κ weight and debiased machine learning estimators that have explicit inverse propensity scores.

A.3. Affine moments

When we verify the conditions of Corollary 5.2 using Auto-κ, we focus on a subclass of the complier parameters in Definition 2.1. This subclass is rich enough to include several empirically important parameters, yet simple enough to avoid iterative estimation. The subclass consists of complier parameters with affine moments, which we now define. The affine moment condition can be relaxed, but doing so incurs iterative estimation (Chernozhukov et al., 2022).

DEFINITION A.1 (AFFINE MOMENT). *We say a doubly robust moment function ψ is affine in θ if it takes the form*

$$\psi(W, \gamma, \alpha, \theta) = A(\theta)\{\gamma(1, X) - \gamma(0, X)\} + \alpha(Z, X)A(\theta)\{V - \gamma(Z, X)\},$$

where $A(\theta)$ is a matrix with entries that are ones, zeros, or components of θ.

We verify that several empirically important complier parameters have affine moments.

DEFINITION A.2 (EMPIRICALLY IMPORTANT COMPLIER PARAMETERS). *Consider the following parameters.*

(1) *LATE is $\theta_0 = E\{Y^{(1)} - Y^{(0)} \mid D^{(1)} > D^{(0)}\}$.*
(2) *Average complier characteristics are $\theta_0 = E\{f(X) \mid D^{(1)} > D^{(0)}\}$ for any measurable function f of covariate X that may have a finite dimensional, real vector value such that $E\{f_j(X)^2\} < \infty$.*
(3) *Complier counterfactual outcome distributions are $\theta_0 = (\theta_0^y)_{y \in \mathcal{U}}$ where*

$$\theta_0^y = \begin{pmatrix} \beta_0^y \\ \delta_0^y \end{pmatrix} = \begin{bmatrix} \mathrm{pr}\{Y^{(0)} \leq y \mid D^{(1)} > D^{(0)}\} \\ \mathrm{pr}\{Y^{(1)} \leq y \mid D^{(1)} > D^{(0)}\} \end{bmatrix}$$

and $\mathcal{U} \subset \mathcal{Y}$ is a fixed grid of finite dimension.

COROLLARY A.1 (EMPIRICALLY IMPORTANT PARAMETERS HAVE AFFINE MOMENTS). *Under Assumption 2.1, the doubly robust moment functions for LATE, average complier characteristics, and complier counterfactual outcome distributions are affine, where*

(1) For LATE (Tan, 2006), we set $V = (Y, D)^\top$ and $A(\theta) = \big(1,\ -\theta\big)$.
(2) For complier characteristics, we set $V = (Df(X)^\top, D)^\top$ and $A(\theta) = \big(I,\ -\theta\big)$.
(3) For complier counterfactual distributions (Belloni et al., 2017), we set

$$V^y = \{(D-1)1_{Y\le y}, D1_{Y\le y}, D\}^\top \text{ and } A(\theta^y) = \begin{pmatrix} 1 & 0 & -\beta^y \\ 0 & 1 & -\delta^y \end{pmatrix}.$$

Proof: Suppose we can decompose $v(w,\theta) = h(w,\theta) + a(\theta)$ for some function $a(\cdot)$ that does not depend on data. Then we can replace $v(w,\theta)$ with $h(w,\theta)$ without changing m and ϕ in the sense of Theorem 4.1. This is because

$$E\{v(W,\theta) \mid z, x\} = E\{h(W,\theta) \mid z, x\} + a(\theta)$$

and hence

$$v(w,\theta) - E\{v(W,\theta) \mid z, x\} = h(w,\theta) - E\{h(W,\theta) \mid z, x\}.$$

Whenever we use this reasoning, we write $v(w,\theta) \propto h(w,\theta)$.

(1) For LATE we can write $\theta_0 = \delta_0 - \beta_0$, where δ_0 is defined by the moment condition $E\{Y^{(1)} - \delta_0 \mid D^{(1)} > D^{(0)}\} = 0$ and β_0 is defined by the moment condition $E\{Y^{(0)} - \beta_0 \mid D^{(1)} > D^{(0)}\} = 0$. Applying Case 2 of Theorem 4.1 to δ_0, we have $v(w,\delta) = d(y-\delta)$. Applying Case 1 of Theorem 4.1 to β_0, we have $v(w,\beta) = (d-1)(y-\beta) \propto (d-1)y - d\beta$. Writing $\theta = \delta - \beta$, the moment function for θ_0 can be derived with

$$v(w,\theta) = v(w,\delta) - v(w,\beta) = y - d\theta.$$

This expression decomposes into $V = (Y, D)^\top$ and $A(\theta) = \big(1,\ -\theta\big)$ in Corollary A.1.

(2) For average complier characteristics, θ_0 is defined by the moment condition $E\{f(X) - \theta_0 \mid D^{(1)} > D^{(0)}\} = 0$. Applying Case 2 of Theorem 4.1 setting $g\{Y^{(1)}, X, \theta_0\} = f(X) - \theta_0$, we have $v(w,\theta) = d\{f(x) - \theta\}$. This expression decomposes into $V = \{Df(X)^\top, D\}^\top$ and $A(\theta) = \big(I,\ -\theta\big)$ in Corollary A.1.

(3) For the complier distribution of $Y^{(0)}$, $\beta_0^{\bar{y}}$ is defined by the moment condition $E\{1_{Y^{(0)}\le\bar{y}} - \beta_0^{\bar{y}} \mid D^{(1)} > D^{(0)}\} = 0$. Applying Case 1 of Theorem 4.1 to $\beta_0^{\bar{y}}$, we have $v(w,\beta^{\bar{y}}) = (d-1)(1_{y\le\bar{y}} - \beta^{\bar{y}}) \propto (d-1)1_{y\le\bar{y}} - d\beta^{\bar{y}}$. For the complier distribution of $Y^{(1)}$, $\delta_0^{\bar{y}}$ is defined by the moment condition $E\{1_{Y^{(1)}\le\bar{y}} - \delta_0^{\bar{y}} \mid D^{(1)} > D^{(0)}\} = 0$. Applying Case 2 of Theorem 4.1 to δ_0, we have $v(w,\delta^{\bar{y}}) = d(1_{y\le\bar{y}} - \delta^{\bar{y}})$. Concatenating $v(w,\beta^{\bar{y}})$ and $v(w,\delta^{\bar{y}})$, we arrive at the decomposition in Corollary A.1.

□

A.4. Inference

We prove the Auto-κ estimator for complier parameters is consistent, asymptotically normal, and semiparametrically efficient. In doing so, we verify the conditions of Corollary 5.2. We build on the theoretical foundations in Chernozhukov et al. (2022) to generalise the main result in Chernozhukov et al. (2022a). We assume the following.

ASSUMPTION A.1 (CONDITIONS FOR COMPLIER PARAMETER ESTIMATION). *Assume*

(1) Affine moment: ψ *is affine in* θ*;*
(2) Bounded propensity: $\pi_0(X)$ *is in* $(\bar{c}, 1-\bar{c})$ *for some* $\bar{c} > 0$ *uniformly over the support of* X*;*
(3) Bounded variance: $\mathrm{var}(V \mid Z, X)$ *is bounded uniformly over the support of* (Z, X)*;*
(4) Nonsingular Jacobian: $J = E\left\{\partial\psi(W, \gamma_0, \alpha_0, \theta)/\partial\theta|_{\theta=\theta_0}\right\}$ *is nonsingular;*
(5) Compact parameter space: $\theta_0, \hat{\theta}$ *are in* Θ*, a compact parameter space;*
(6) Rates: $|\hat{\alpha}|_\infty = O_p(1)$*,* $\|\hat{\alpha} - \alpha_0\| = o_p(1)$*,* $\|\hat{\gamma} - \gamma_0\| = o_p(1)$*, and* $\|\hat{\alpha} - \alpha_0\|\|\hat{\gamma} - \gamma_0\| = o_p(n^{-1/2})$*, where* $\|V_j\| = \{E(V_j^2)\}^{1/2}$ *and* $\|V\| = \{\|V_1\|, \ldots, \|V_{dim(V)}\|\}^\top$*.*

The most substantial condition in Assumption A.1 is the rate condition. In Online Appendix S1, we verify the rate condition for the $\hat{\alpha}$ estimator in Algorithm A.2. Since $\hat{\gamma}$ is a standard nonparametric regression, a broad variety of estimators and their mean square rates can be quoted to satisfy the rate condition for $\hat{\gamma}$. The product condition formalises the mixed bias property. It allows *either* the convergence rate of $\hat{\gamma}$ to be slower than $n^{-1/4}$ *or* the convergence rate of $\hat{\alpha}$ to be slower than $n^{-1/4}$, as long as the other convergence rate is faster than $n^{-1/4}$. As such, it allows *either* $\hat{\gamma}$ to be a complicated function *or* $\hat{\alpha}$ to be a complicated function, as long as the other is a simple function, in a sense that we formalise in Online Appendix S1.

THEOREM A.1 (CONSISTENCY AND ASYMPTOTIC NORMALITY). *Suppose Assumption A.1 holds. Then* $n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, C)$ *and* $\hat{C} = C + o_p(1)$ *where*

$$J = E\left\{\frac{\partial\psi_0(W)}{\partial\theta}\right\}, \ \hat{J} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\frac{\partial\hat{\psi}_i(\hat{\theta})}{\partial\theta}, \ \Omega = E\{\psi_0(W)\psi_0(W)^\top\}, \ \hat{\Omega} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i(\hat{\theta})\hat{\psi}_i(\hat{\theta})^\top$$

$$C = J^{-1}\Omega J^{-1}, \quad \hat{C} = \hat{J}^{-1}\hat{\Omega}\hat{J}^{-1}, \quad \psi_0(W) = \psi(W, \gamma_0, \alpha_0, \theta_0), \quad \hat{\psi}_i(\theta) = \psi(W_i, \hat{\gamma}_{-\ell}, \hat{\alpha}_{-\ell}, \theta).$$

Proof: We defer the proof to Online Appendix S2. □

When the doubly robust moment function ψ is affine in θ, the Auto-κ estimator achieves semi-parametric efficiency because the doubly robust moment function coincides with the semi-parametrically efficient score (Hahn, 1998). Therefore hypothesis tests based on Auto-κ are asymptotically efficient in this case. When the doubly robust moment function ψ is not affine in θ, the Auto-κ estimator may not be semi-parametrically efficient, and so hypothesis tests based on Auto-κ may not be asymptotically efficient. Future research may examine the power properties of tests based on Auto-κ when ψ is not affine in θ.

Throughout this paper, we focus on low dimensional complier parameters identified using a binary instrument Z, which is valid conditional on a possibly high dimensional vector of covariates X. Future work may consider high dimensional complier parameters, e.g., complier counterfactual outcome distributions or complier characteristic distributions using a grid of increasing dimension. When the grid has fixed dimension, then the complier parameters are low dimensional and so our inference and efficiency results apply.

In summary, extensions of our Auto-κ inference and efficiency results to non-affine and high dimensional complier parameters are important directions for future work.

APPENDIX B: EXTENSIONS: VARIANCES AND DISTRIBUTIONS

B.1. Scope of extensions

As discussed in Section 1, the focus of this paper is low dimensional complier parameters that are identified using a binary instrumental variable Z, which is valid conditional on a possibly high dimensional vector of covariates X. As defined in Section 2, the average complier characteristics belong to this class, where $\theta_0 = E\{f(X) \mid D^{(1)} > D^{(0)}\}$ for a function f of covariate X that has a finite dimensional, real vector value.

In this Appendix, we demonstrate that specific choices of f allow us to extend our results to complier characteristic variances and distributions. Formally, our framework extends to test (a) a finite number of moments of complier characteristics and (b) a finitely supported distribution of complier characteristics. For simplicity, we state these extensions for a scalar characteristic of interest, which we denote by $X_* \subset X$. These extensions generalise to vector characteristics of fixed dimension, with heavier notation.

The inference results of Appendix A apply to parameter vectors of fixed dimension. When covariates are high dimensional, we test means of 'few' covariates (i.e., finitely many covariates), or test distributions of 'simple' covariates (i.e., covariates with finite support). The results in Appendix A do not apply to complier characteristics of increasing dimension, such as means of 'many' covariates (e.g., all of the high dimensional covariates) or distributions of 'complex' covariates (i.e., covariates with increasing support). Such extensions are important directions for future work.

B.2. Corollaries for complier characteristic variances and distributions

Suppose we wish to test the null hypothesis that two different instruments Z_1 and Z_2 induce complier subpopulations with the same variances of the observable characteristic X_*, which is a scalar covariate. Set $f(X) = (X_*, X_*^2)^\top$, so that

$$(\theta_F, \theta_S)^\top = E\{(X_*, X_*^2)^\top \mid D^{(1)} > D^{(0)}\}.$$

θ_F is the first moment and θ_S is the second moment. Denote by $(\hat{\theta}_{1F}, \hat{\theta}_{1S})$ and $(\hat{\theta}_{2F}, \hat{\theta}_{2S})$ the estimators for these moments using the different instruments Z_1 and Z_2, respectively. The following procedure allows us to test the null hypothesis from some point estimator $\hat{\theta} = (\hat{\theta}_{1F}, \hat{\theta}_{1S}, \hat{\theta}_{2F}, \hat{\theta}_{2S})^\top$ and from some variance estimator $\hat{C}$ for the asymptotic variance of $\hat{\theta}$. Appendix A provides details for constructing $\hat{\theta}$ and $\hat{C}$ based on the Auto-κ approach.

ALGORITHM B.1 (TEST FOR DIFFERENCE OF COMPLIER CHARACTERISTIC VARIANCES). *Given $\hat{\theta}$ and $\hat{C}$, which may be based on Auto-κ as in Appendix A,*

STEP *1. Calculate the statistic $T = n\{\hat{\theta}_{1S} - \hat{\theta}_{1F}^2 - (\hat{\theta}_{2S} - \hat{\theta}_{2F}^2)\}(R\hat{C}R^\top)^{-1}\{\hat{\theta}_{1S} - \hat{\theta}_{1F}^2 - (\hat{\theta}_{2S} - \hat{\theta}_{2F}^2)\}$ where $R = \left(-2\hat{\theta}_{1F},\ 1,\ 2\hat{\theta}_{2F}, -1\right)$.*

STEP *2. Compute the value c_a as the $(1-a)$ quantile of $\chi^2(1)$.*

STEP *3. Reject the null hypothesis if $T > c_a$.*

COROLLARY B.1 (TEST FOR DIFFERENCE OF COMPLIER CHARACTERISTIC VARIANCES). *If $n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, C)$ and $\hat{C} = C + o_p(1)$, then the hypothesis test in Algorithm B.1 falsely rejects the null hypothesis H_0 with probability approaching the nominal level, i.e., $\mathrm{pr}(T > c_a \mid H_0) \to a$.*

Proof: The result is immediate from Newey and McFadden (1994, sect. 9). The argument is identical to that of Corollary 5.2, with further appeal to the delta method. □

Next, suppose we wish to test the null hypothesis that two different instruments Z_1 and Z_2 induce complier subpopulations with the same distributions of the observable characteristic X_*, which is a scalar covariate. Further suppose that X_* has a finite support of d values, which we denote $\mathcal{U} = (u_1, \ldots, u_d)$. Set $f(X) = (1_{X_* \le u_1}, \ldots, 1_{X_* \le u_d})^\top$, so that

$$(\theta^{u_1}, \ldots, \theta^{u_d})^\top = E\{(1_{X_* \le u_1}, \ldots, 1_{X_* \le u_d})^\top \mid D^{(1)} > D^{(0)}\}.$$

In this notation, $\theta^u = \mathrm{pr}\{X_* \le u \mid D^{(1)} > D^{(0)}\}$ is the cumulative mass function of the complier characteristic X_* evaluated at value $u \in \mathcal{U}$. Denote by $\hat{\theta}_1 = (\hat{\theta}_1^{u_1}, \ldots, \hat{\theta}_1^{u_d})^\top$ and $\hat{\theta}_2 = (\hat{\theta}_2^{u_1}, \ldots, \hat{\theta}_2^{u_d})^\top$ the estimators for these cumulative mass functions using the different instruments Z_1 and Z_2, respectively. The following procedure allows us to test the null hypothesis from some point estimator $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$ and from some variance estimator $\hat{C}$ for the asymptotic variance of $\hat{\theta}$. Appendix A provides details for constructing $\hat{\theta}$ and $\hat{C}$ based on the Auto-κ approach.

Table B1. Comparison of complier characteristic distributions.

Mother's schooling category	Twins	Same sex	2 sided
High school dropout	0.16	0.19	0.13
(SE)	(0.01)	(0.02)	–
High school graduate	0.47	0.54	<0.01
(SE)	(0.02)	(0.02)	–
Some college	0.22	0.19	0.06
(SE)	(0.01)	(0.01)	–
College graduate	0.13	0.08	<0.01
(SE)	(0.01)	(0.01)	–

Notes: SE, standard error. See Appendix A and Online Appendix S5 for estimation details.

ALGORITHM B.2 (TEST FOR DIFFERENCE OF COMPLIER CHARACTERISTIC DISTRIBUTIONS). *Given $\hat{\theta}$ and $\hat{C}$, which may be based on Auto-κ as in Appendix A,*

STEP *1. Calculate the statistic $T = n(\hat{\theta}_1 - \hat{\theta}_2)^\top (R\hat{C}R^\top)^{-1}(\hat{\theta}_1 - \hat{\theta}_2)$ where $R = \big(I, -I\big)$.*
STEP *2. Compute the value c_a as the $(1-a)$ quantile of $\chi^2(d)$.*
STEP *3. Reject the null hypothesis if $T > c_a$.*

COROLLARY B.2 (TEST FOR DIFFERENCE OF COMPLIER CHARACTERISTIC DISTRIBUTIONS). *If $n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, C)$ and $\hat{C} = C + o_p(1)$, then the hypothesis test in Algorithm B.2 falsely rejects the null hypothesis H_0 with probability approaching the nominal level, i.e.,* $\mathrm{pr}(T > c_a \mid H_0) \to a$.

Proof: The result is immediate from Newey and McFadden (1994, sect. 9). □

In summary, for appropriate choices of f in Definition 2.2, our results extend from complier characteristic averages to complier characteristic variances and complier characteristic distributions over finite support. These additional hypothesis tests follow directly from the results in the main text. The extension of these results to high dimensional complier parameters, e.g., an increasing number of moments or an increasing support $\mathcal{U}$, is an important direction for future work.

B.3. Empirical application

Finally, we implement our generalised hypothesis test to evaluate whether two different instruments induce different distributions of complier characteristics. In the empirical application of Section 5, mother's schooling may be discretised as a categorical random variable that takes on four values: high school dropout, high school graduate, some college, and college graduate. In what follows, we set X_* to be the mother's schooling, and we set $\mathcal{U} = (u_1, \ldots, u_4)$ to be these four categories of schooling.

Table B1 summarises results. In Columns 2 and 3, we present probability mass function estimates; taking sums recovers cumulative mass function estimates. Columns 4 reports the p-values for tests of individual probability mass function values: for each category of mother's schooling, we test the null hypothesis that the probability mass function value is equal for the twins and same-sex instruments. We find strong evidence of a difference for the highest education category, i.e., for college graduates.

Next, we conduct a test of the null hypothesis that the probability mass function is equal across all categories of mother's schooling, for the twins and same-sex instruments. The p-value of this joint test is less than 0.01. In summary, we find evidence of a difference in distributions of mother's education for twins and same-sex compliers, likely due to a difference in the highest education category.