MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

A test for instrumental variable validity using a correlation restriction

Discussion Paper no. 2025-06

Ratbek Dzhumashev and Ainura Tursunalieva

Abstract:

DiTraglia and García-Jimeno (2021) demonstrate that the correlation coefficients between an IV, an endogenous regressor, and the outcome variable must satisfy a specific joint constraint determined by their relationships with the structural error term. We exploit this constraint to develop a novel Correlation Restriction test that becomes feasible when the direction of endogeneity bias is known. Our test quantifies the probability of instrument orthogonality to the structural error across the plausible range of endogeneity magnitudes, providing researchers with a previously unavailable diagnostic tool in the frequentist setting. Through simulations and applications to diverse empirical settings including returns to education, criminal recidivism, and development economics, we establish that our method reliably identifies invalid instruments and characterizes the endogeneity range over which valid instruments maintain their exogeneity. This approach contributes to instrumental variable methods by transforming a key identification assumption from an untestable assertion into an empirically verifiable condition.

Keywords: endogeneity, validity of instrumental variable, linear regression

JEL Classification: C18, C26, C36, C52

Ratbek Dzhumashev: Department of Economics, Monash University (email: Ratbek.Dzhumashev@monash.edu); Ainura Tursunalieva: Data61, CSIRO (email: ainura.tursunalieva@data61.csiro.au).

A test for instrumental variable validity using a correlation restriction

Ratbek Dzhumashev*and Ainura Tursunalieva†

February 24, 2025

Abstract

DiTraglia and García-Jimeno (2021) demonstrate that the correlation coefficients between an IV, an endogenous regressor, and the outcome variable must satisfy a specific joint constraint determined by their relationships with the structural error term. We exploit this constraint to develop a novel Correlation Restriction test that becomes feasible when the direction of endogeneity bias is known. Our test quantifies the probability of instrument orthogonality to the structural error across the plausible range of endogeneity magnitudes, providing researchers with a previously unavailable diagnostic tool in the frequentist setting. Through simulations and applications to diverse empirical settings including returns to education, criminal recidivism, and development economics, we establish that our method reliably identifies invalid instruments and characterizes the endogeneity range over which valid instruments maintain their exogeneity. This approach contributes to instrumental variable methods by transforming a key identification assumption from an untestable assertion into an empirically verifiable condition.

Keywords: endogeneity, validity of instrumental variable, linear regression
JEL codes: C18, C26, C36, C52

1 Introduction

Instrumental variable (IV) estimation represents a cornerstone methodology for identifying causal effects in the presence of endogeneity. The validity of this approach hinges on two fundamental conditions: exogeneity - the instrument must be uncorrelated with the structural error term ($E(z'u) = 0$) - and relevance - the instrument must sufficiently correlate with the endogenous regressor. While relevance can be empirically verified through first-stage diagnostics, the exogeneity condition remains inherently untestable in standard frameworks because the structural error is unobserved. This limitation has led many researchers to rely predominantly on theoretical arguments, institutional knowledge, and informal reasoning to justify instrument validity.

*Department of Economics, Monash University, Ratbek.Dzhumashev@monash.edu
†Data61, CSIRO, ainura.tursunalieva@data61.csiro.au

The fundamental challenge of testing instrument validity has been formally articulated by Pearl (1995), who conjectured that with continuously distributed endogenous variables, instrument validity becomes theoretically untestable. Recent theoretical advancements by Gunsilius (2021) have demonstrated that imposing weak structural assumptions can partially restore testability in specific contexts. This paper complements this literature by establishing that instrument validity becomes systematically testable even with continuous endogenous regressors when sign restrictions on the endogeneity correlation are incorporated into the analysis of correlation restrictions. Our approach transforms a previously untestable condition into an empirically verifiable criterion, providing researchers with a practical method to evaluate instrument exogeneity when the direction of endogeneity bias is known.

Building on DiTraglia and García-Jimeno (2021), we exploit the relationship between structural and reduced-form correlations. These authors demonstrate that the correlation structure among the instrument, endogenous regressor x, outcome variable y, and unobserved structural error u must necessarily adhere to the following constraint:

$$\rho_{uz} = \rho_{xz}\rho_{xu} - (\rho_{xy}\rho_{xz} - \rho_{yz})\sqrt{\frac{1-\rho_{xu}^2}{1-\rho_{xy}^2}},$$

where ρ_{ij} denotes the correlation coefficient between variables i and j, $corr(i, j)$.

This relationship, combined with sign restrictions on $corr(x, u)$, enables us to develop a novel test for IV validity. Notably, sign restrictions on $corr(x, u)$ are plausibly available in many empirical contexts where IV estimation is employed. Researchers frequently possess domain-specific knowledge about the direction of endogeneity bias, even when its precise magnitude remains unknown (Moon and Schorfheide, 2009). DiTraglia and García-Jimeno (2021) incorporate such sign and interval restrictions on regressor endogeneity as informative priors to narrow the identified set in their Bayesian framework. In related work, Nevo and Rosen (2012) establish partial identification results by explicitly connecting the sign of $cov(x, u)$ to structural parameters, thereby deriving bounds for causal effects when instruments are potentially invalid.

In our context, when the sign of $corr(x, u)$ is known, we can determine whether the range of possible values for $\hat{\rho}_{uz}$ excludes zero, thereby testing IV validity. Specifically, when a range of correlation is given by $\rho_{xu} \in D$, one can calculate the range of admissible values C for ρ_{zu} substituting $\rho_{xu} \in D$ in the above correlation restriction equation, and check whether $\rho_{zu} \in C$ excludes 0. Intuitively, if admissible

values for ρ_{zu} determined this way do not contain zero, then this implies that $E(z'u) \neq 0$; thus, the given IV cannot be valid.

To implement the above testable implication, we follow the inference procedure suggested DiTraglia and García-Jimeno (2021, see the supplement) that explains the conditions under which the Bayesian posterior and frequentist large-sample distributions are compatible. This allows us to appeal to Theorem 5 of Kline & Tamer (2016) to show that a credible set for the identified parameters can also be an exact point-wise Frequentist confidence set.[1] In this light, we develop a novel Correlation Restriction (CR) test using advanced bootstrapping techniques (Davidson and Hinkley, 2009; Efron and Tibshirani, 1994) to evaluate whether the estimated correlation between the instrument and structural error ($\hat{\rho}_{uz} = 0$) lies within the feasible range implied by the estimated endogeneity correlation ($\hat{\rho}_{xu}$). Our approach establishes a dual criterion for instrument validity: (1) the coverage probability $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$ must be strictly positive, and (2) the conventional 95% confidence interval for $\hat{\rho}_{uz}$ must contain zero, satisfying standard exogeneity requirements.

Our comprehensive empirical evaluation demonstrates the CR test's diagnostic power across diverse settings. In controlled simulation environments, we establish that instruments failing our CR test systematically produce biased parameter estimates, validating its discriminatory capability. When applied to Card's (1993) seminal study on returns to education, our test reveals that two instrumental variables previously employed fail the CR criteria, explaining their documented contribution to OLS bias, while two other instruments technically pass but exhibit low coverage probabilities, signalling limited reliability. We further validate the test's robustness through applications across four additional empirical studies spanning different economic contexts, confirming its effectiveness in detecting invalid instruments under varying data structures and model specifications.

Our paper makes three contributions to the literature on instrumental variable validity assessment: First, we extend the joint correlation restrictions framework established by DiTraglia and García-Jimeno (2021) by developing a practical frequentist inference procedure that enables testing IV validity when the sign of $corr(x,u)$ is known. Second, we conduct comprehensive evaluations using both sim-

[1] See also Kitagawa (2012) and Chen et al. (2018), who provide sufficient conditions under which such inferences obtained in the Bayesian framework have a valid frequentist interpretation.

ulated and empirical data, demonstrating the test's robustness to violations of normality and linearity assumptions that often arise in applied settings. Third, we advance the theoretical understanding of IV testability by showing how incorporating sign restrictions on endogeneity can restore testability in settings with continuous endogenous regressors, where traditional approaches are typically uninformative.

This study contributes to the evolving literature on testing and identifying assumptions in econometric models. The foundational work of Pearl (1995) introduced instrumental inequality, a theoretical constraint for instrument validity that necessitated discretization of the endogenous treatment, while conjecturing that validity testing becomes infeasible with continuous treatments. The literature has since evolved along multiple trajectories. Bonet (2001) provided formal proof of Pearl's conjecture for discrete outcomes, while Manski (2003) established an equivalent inequality in the context of missing data models. Kitagawa (2015) extended this framework by developing a formal test for continuous outcomes with binary treatments and instruments. Wang et al. (2017) further contributed empirical testing procedures specifically for binary data structures. Kédagni and Mourifié (2020) offered comprehensive confirmation of Pearl's conjecture for binary variables, generalizing the results to accommodate discrete instruments with arbitrary outcome distributions. Meanwhile, Jiang & Ding (2020) derived identification bounds for binary models incorporating measurement error, and Acerenza et al. (2023) applied analogous methodology to test identifying assumptions within bivariate probit specifications. A parallel line of inquiry has emerged focusing on the evaluation of identifying assumptions by directly testing exclusion restrictions without relying on conventional orthogonality conditions (Kiviet, 2020; D'Haultfœuille et al., 2021).

Despite these substantial theoretical advancements, the testability of instrument validity in models with continuous endogenous variables - a frequent case in applied econometrics - remains notably underdeveloped. A significant exception is Gunsilius (2021), who demonstrated that imposing weaker structural assumptions can enhance testability properties. Recent methodological innovations have primarily gravitated toward machine-learning frameworks (Xie et al., 2022, 2020; Silva & Shimizu, 2017), which do not advance conventional econometric approaches accessible to applied researchers.

Our paper addresses this critical gap by developing a rigorous test for instrument validity anchored

in orthogonality conditions for continuous variable settings. The methodological innovation lies in our leveraging of the joint correlation restriction connecting instrument invalidity, treatment endogeneity, and the structural error term - as established by DiTraglia and García-Jimeno (2021) - combined with plausible sign restrictions on the endogeneity correlation $corr(x,u)$. This approach yields a practical testing procedure that can be readily implemented using standard econometric methods.

The remainder of the paper is structured as follows: In Section 2, we outline the relationship between the reduced-form coefficients and the structural equation parameters, which we use to establish the CR test for IVs, provided the known sign of $corr(x,u)$. In Section 3, we present a numerical illustration of the CR test and its application to the IV method. In Section 4, we conclude the study.

2 Model

Consider a linear model with endogeneity, expressed as a triangular system:

$$\tilde{y} = \tilde{x}'\beta + H'\gamma + u, \tag{1}$$

$$\tilde{x} = \tilde{z}'\pi + H'\eta + v, \tag{2}$$

where $\tilde{x}$ is a potentially endogenous regressor, $\tilde{y}$ is an outcome of variable, and H is a matrix of exogenous controls including the constant. All variables are assumed to be scalar random variables. Endogeneity arises from $E(\tilde{x},u) \neq 0$, which may result from omitted variables, measurement errors, or simultaneity. The model estimates become biased when the IV $\tilde{z}$ is not orthogonal to the structural error u. To develop our correlation restriction (CR) test, we first express the model in terms of reduced-form variables. Define:

$$y = \tilde{y} - H'\alpha_y,\ x = \tilde{x} - H'\alpha_x, \text{ and } z = \tilde{z} - H'\alpha_z, \tag{3}$$

where $(\alpha_y, \alpha_x, \alpha_z) \equiv A = E[HH']^{-1}E[H'\tilde{w}]$ for $\tilde{w} \in \{\tilde{y},\tilde{x},\tilde{z}\}$. This transformation yields the reduced-form system:

$$y = \beta x + u, \tag{4}$$

$$x = \pi z + \varepsilon. \tag{5}$$

Following the theoretical framework established by DiTraglia and García-Jimeno (2021), we formalize our analysis with the following assumptions:

Assumption 1 (Model structure). The researcher observes $(\tilde{y},\ \tilde{x},\ \tilde{z})$. Recall that $(y,\ x,\ z)$ are generated from (3), where (i) $cov(z,v)=0$; (ii) z is relevant for x: $\pi \neq 0$; (iii) H includes all exogenous variables and the constant term so that $E[u]=E[\varepsilon]=0$; (iv) x is correlated with u : $cov(x,u)\neq 0$.

Assumption 2.(Covariance structure). The covariance matrix Ω of (z,u,ε) exists and is positive definite. This matrix is given by the following equations:

$$\Omega = \begin{bmatrix} \sigma_u^2 & \sigma_{u\varepsilon} & \sigma_{uz} \\ \sigma_{u\varepsilon} & \sigma_\varepsilon^2 & 0 \\ \sigma_{uz} & 0 & \sigma_z^2 \end{bmatrix}.$$

Under these assumptions, DiTraglia and García-Jimeno (2021) state the key theoretical result that underpins our testing procedure:

Theorem 2.1 *Under Assumptions 1 and 2, the following equality holds:*

$$\rho_{uz} = \rho_{xz}\rho_{xu} - (\rho_{xy}\rho_{xz} - \rho_{yz})\sqrt{\frac{1-\rho_{xu}^2}{1-\rho_{xy}^2}}, \tag{6}$$

where ρ_{ij} denotes the correlation coefficient between variables i and j.

Proof (refer to the proof of Proposition 2.1. in DiTraglia and García-Jimeno (2021))

This theorem establishes a fundamental relationship between regressor endogeneity ($\rho_{xu}\neq 0$) and instrument invalidity ($\rho_{uz}=0$), mediated by the observable correlations in the data. It provides the theoretical foundation for our CR test by expressing ρ_{uz} as an explicit function of ρ_{xu} and observable correlations.

We develop a test for IV exogeneity based on the correlation restriction that must be satisfied when z is a valid instrument, given knowledge of $sign[corr(x,u)]$. This test exploits the relationship established in Theorem 2.1, which links observable correlations to unobservable relationships between the instrument and structural error. The key insight from Theorem 2.1 is that the equation applied to data

$$\hat{\rho}_{uz} = \hat{\rho}_{xz}\hat{\rho}_{xu} - (\hat{\rho}_{xy}\hat{\rho}_{xz} - \hat{\rho}_{yz})\sqrt{\frac{1-\hat{\rho}_{xu}^2}{1-\hat{\rho}_{xy}^2}} \tag{7}$$

connects the unobservable correlations ρ_{xu} and ρ_{uz} with estimable correlation coefficients. When we know the sign of ρ_{xu}, we can leverage this relationship to test IV validity. For instance, if $\rho_{xu} > 0$, we can compute the range of admissible values for $\hat{\rho}_{uz}$ over $\hat{\rho}_{xu} \in (0,1)$ and check whether this range includes zero. This intuition leads to our main testable implication:

Corollary 2.2 *Let (x,y,z) satisfy Assumptions 1 and 2. Define the mapping $\hat{\rho}_{uz} : D \to C$, given as*

$$\hat{\rho}_{uz} = \hat{\rho}_{xz}\hat{\rho}_{xu} - (\hat{\rho}_{xy}\hat{\rho}_{xz} - \hat{\rho}_{yz})\sqrt{\frac{1-\hat{\rho}_{xu}^2}{1-\hat{\rho}_{xy}^2}},$$

where C is the range of values for $\hat{\rho}_{uz}$, and D is the domain for $\hat{\rho}_{xu}$, defined as:

$$D = \begin{cases} (0,1) & \text{if } corr(x,u) > 0 \\ (-1,0) & \text{if } corr(x,u) < 0 \end{cases}$$

If z is a valid IV, then $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D) > 0$.

Proof The data for x, y, and z allow to determine the estimates $\hat{\rho}_{xy}$, $\hat{\rho}_{xz}$, and $\hat{\rho}_{yz}$. Given that $\rho_{xu} \equiv cor(x,u) \neq 0$, we have $\hat{\rho}_{xu} \in D$. Then, $\hat{\rho}_{uz} = \hat{\rho}_{xz}\hat{\rho}_{xu} - (\hat{\rho}_{xy}\hat{\rho}_{xz} - \hat{\rho}_{yz})\sqrt{\frac{1-\hat{\rho}_{xu}^2}{1-\hat{\rho}_{xy}^2}} \in C$ is determined on $\hat{\rho}_{xu} \in D$. Given that $sign[corr(x,u)]$ is the true sign of ρ_{xu}, the true ρ_{xu} is contained in D. Then, $\hat{\rho}_{uz} : D \to C$ implies that for a valid IV satisfying $\rho_{uz} = 0$, its prediction is found as $E[\hat{\rho}_{uz}] = 0 \in C$. Therefore, $0 \in [min(\hat{\rho}_{uz}), max(\hat{\rho}_{uz})$, thus, $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D) > 0$ holds for a valid IV, z.■

To implement this result empirically, we develop a bootstrap-based testing procedure:

- Generate bootstrap samples to estimate the sampling distribution of the correlation coefficients.
- For each sample, compute $\hat{\rho}_{uz}$ across the domain D of possible $\hat{\rho}_{xu}$ values.
- Calculate two key statistics across all bootstraps:
  - Coverage probability, $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$, and
  - 95% confidence intervals for $\hat{\rho}_{uz}$.

Determining a 95% CI may seem superfluous, given if $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D) > 0$, it should contain 0. However, by adjusting the range of the endogeneity ρ_{xu}, we can determine its narrowest

interval, where $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$ is the same as for the maximum interval of D given by $(0,1)$ or $(-1,0)$ and the CI contains $\hat{\rho}_{uz} = 0$. For example, we one uses too wide range for $\hat{\rho}_{xu}$ and finds that $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D) > 0$. However, 95% CI may not contain 0. This outcome indicates that $\hat{\rho}_{xu}$ is feasible in a narrower range than it is assumed.

This leads to our formal test definition:

Correlation restriction test for IVs: Given the feasible range of values for ρ_{xu}, a valid IV satisfies the following two conditions: (1) the coverage probability $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$ is positive and (2) 95% CI of $\hat{\rho}_{uz}$ contains 0.

Important Caveats: The CR test's reliability depends crucially on: (i) Correct specification of $sign[corr(x,u)]$, (ii) Validity of the underlying model assumptions, (iii) Appropriate functional form specification. Test failure may indicate either IV invalidity or violation of these maintained assumptions. Therefore, we recommend conducting thorough specification tests before applying the CR test.

2.1 Asymptotic distribution of sample correlation coefficients

The statistical properties of sample correlation coefficients are crucial for our testing procedure. For any two variables (x,y), the sample correlation coefficient is defined as:

$$\hat{\rho}xy = \frac{sxy}{s_x s_y},$$

where s_{xy} denotes the sample covariance and (s_x, s_y) are the sample standard deviations. Under bivariate normality, the asymptotic distribution of the sample correlation coefficient is well-established (Lehmann, 2010, ch 5.4):

$$\hat{\rho}_{xy} \overset{a}{\sim} N(\rho_{xy}, \frac{(1-\rho_{xy}^2)^2}{n}),$$

where ρ_{xy} is the population correlation coefficient and n is the sample size. However, the variance of this distribution depends on the unknown parameter ρ_{xy}, complicating inference. To address this, Fisher (1921) introduced the transformation:

$$\hat{\rho}'_{xy} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{xy}}{1-\hat{\rho}_{xy}}\right),$$

which yields the simpler asymptotic distribution:

$$\sqrt{n}(\hat{\rho}'_{xy} - \rho'_{xy}) \xrightarrow{d} N(0,1).$$

The consistency and asymptotic normality of $\hat{\rho}_{xy}$ follow from two fundamental theorems: The Continuous Mapping Theorem ensures that $\hat{\rho}_{xy} \xrightarrow{p} \rho_{xy}$ since the correlation coefficient is a continuous function of sample moments. The Glivenko-Cantelli Theorem guarantees uniform convergence of empirical distributions to their population counterparts (van der Vaart, 1998, p. 265), establishing the consistency of our correlation estimates. These asymptotic properties provide the theoretical foundation for the statistical inference procedures used in our CR test.

2.2 Empirical implementation

The implementation of our CR test relies on bootstrap methods to estimate both the coverage probability of $\hat{\rho}_{uz} = 0$ and its confidence intervals. The core function from Corollary 2.2 is:

$$\hat{\rho}_{uz} = \hat{\rho}_{xz}\hat{\rho}_{xu} - (\hat{\rho}_{xy}\hat{\rho}_{xz} - \hat{\rho}_{yz})\sqrt{\frac{1-\hat{\rho}_{xu}^2}{1-\hat{\rho}_{xy}^2}}.$$

For fixed $\hat{\rho}_{xu} \in D$, this function's randomness stems solely from the sample correlations $\hat{\rho}_{xz}$ and $\hat{\rho}_{yz}$. We employ multivariate bootstrapping, resampling intact rows from the original dataset S times to preserve the dependence structure among variables (Davidson and Hinkley, 2009; Efron and Tibshirani, 1994). For each bootstrap sample $i = 1, \ldots, S$, we:

(i) Calculate sample correlations $\hat{\rho}_{xy_i}$, $\hat{\rho}_{xz_i}$, and $\hat{\rho}_{yz_i}$,

(ii) Compute $\hat{\rho}_{uz_i}$ across the domain D of $\hat{\rho}_{xu}$,

(iii) Evaluate whether $\hat{\rho}_{uz_i} = 0$ falls within the computed range.

The coverage probability $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$ is then estimated as the proportion of samples satisfying condition (iii). A researcher needs to adjust the interval covered by D and determine its smallest size that renders the same value of the coverage probability for $\hat{\rho}_{uz} = 0$ as D of the maximum feasible size.

For robust inference, we implement three complementary confidence interval approaches (Li, 2022). All confidence intervals are computed using the pooled values of $\hat{\rho}_{uz_i}$ across both bootstrap samples and

the range of $\hat{\rho}_{xu}$ values, ensuring comprehensive coverage of the parameter space. The Basic Bootstrap CI utilizes Fisher's transformation:

$$\hat{\rho}'_{uz} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{uz}}{1-\hat{\rho}_{uz}}\right)$$

with the 95% CI given by:

$$CI_{.95} = \hat{\rho}'_{uz} \pm 1.96 \cdot \left(1/\sqrt{n-3}\right)$$

followed by back-transformation to the correlation scale.

To address potential skewness in the sampling distribution, we also compute the Bootstrap Percentile Interval (BPI):

$$BPI = [\hat{\rho}_{uz}(\alpha/2), \hat{\rho}_{uz}(1-\alpha/2)]$$

where α is the significance level. This approach offers greater robustness by relying on empirical quantiles rather than assuming normality.

For the most comprehensive inference, we employ the Bias-Corrected and Accelerated (BCa) Bootstrap CI (Chan & Chan, 2004; Puth et al., 2015). This method refines the BPI by adjusting for both bias and skewness, providing superior coverage properties when the sampling distribution deviates from normality. The BCa method is particularly valuable for moderate to small sample sizes and when the underlying distribution shows substantial non-normality.

3 A numerical illustration

We demonstrate the performance of the CR test using simulated data where the true data-generating process (DGP) is known. This approach allows us to evaluate the test's ability to identify valid instruments when the sign of $corr(x,u)$ is known.

3.1 Data Generating Process

Our simulation begins with regressor $x \sim U[0,20]$. We then create two variants of the disturbance term:

$$u_p = 0.6x + e \quad \text{and} \quad u_n = -0.6x + e, \quad \text{where } e \sim N(0,10)$$

yielding $corr(x,u_p) = 0.3298$ and $corr(x,u_n) = -0.3268$.[2] The outcome variable is generated as:

$$y_p = 2x + u_p$$

Finally, we construct two instrumental variables: z_p satisfying $corr(z_p,u_p) = 0$ and z_n satisfying $corr(z_n,u_n) = 0$.

Table 1 presents the 99% confidence intervals and means of sample correlation coefficients based on 10,000 samples of 500 observations each. The narrow confidence intervals and consistent signs of correlations ensure reliable inference using sample means.

Table 1: 99%-confidence intervals and means of sample correlation coefficients

	$\hat{\rho}_{yz}$	$\hat{\rho}_{yx}$	$\hat{\rho}_{xz}$
z_p	$0.603 \pm 4 \cdot 10^{-4}$	$0.832 \pm 2 \cdot 10^{-4}$	$0.944 \pm 1 \cdot 10^{-4}$
z_n	$0.787 \pm 3 \cdot 10^{-4}$	$0.832 \pm 2 \cdot 10^{-4}$	$0.945 \pm 1 \cdot 10^{-4}$

z_p satisfies the CR; z_n fails the CR.

$\hat{\rho}_{yz} \equiv corr(y,z)$, $\hat{\rho}_{xz} \equiv corr(x,z)$, and $\hat{\rho}_{xy} \equiv corr(x,y)$.

Distributions based on 10000 samples with 500 observations.

3.2 Test Results

The CR test reveals stark differences between our two instruments. For z_p, we find a coverage probability $\mathbb{P}(\hat{\rho}_{uz} = 0 \in C | \hat{\rho}_{xu} \in D)$ of 0.99 and a 95% CI for $\hat{\rho}_{uz}$ of [-0.024, 0.014], supporting its validity. The corresponding IV estimates yield a 95% CI for βIV of [1.96, 2.04], accurately capturing the true parameter value of 2. In contrast, z_n fails both test criteria with a coverage probability of 0.0 and a 95% CI for $\hat{\rho}_{uz}$ of [0.290, 0.326]. This invalidity manifests in biased IV estimates, with a 95% CI for βIV of [2.27, 2.64] that excludes the true parameter value. These differences are reflected in the point estimates:

$$\hat{\beta} z_p = 2.000 \pm 0.04 \quad \text{vs} \quad \hat{\beta} z_n = 2.609 \pm 0.04 \quad (95\% \text{ CIs}).$$

[2] We used $set.seed(1000)$ for x, $set.seed(1001)$ for u_p and $set.seed(2000)$ for u_n.

Table 2: The correlation restrictions test for z_i, $i \in [p,n]$

	Parameter estimates		*Inference*		
IVs $\tilde{z}$	OLS: $\hat{\beta}$	IV: $\hat{\beta}_{IV}$	$P(\hat{\rho}_{uz}=0)>0$	95% CI:$\hat{\rho}_{uz}$	95% CI:$\hat{\beta}_{IV}$
	2.607^{***} (0.005)				
z_p		2.000^{***} (0.006)	0.99	[-0.024,0.014]	[1.96,2.04]
z_n		2.609^{***} (0.006)	0.0	[0.290,0.326]	[2.27, 2.64]

By construction $\rho_{xu} = 0.3298$.

$P(\rho_{uz}=0)>0$ *gives the fraction of draws compatible with a valid instrument* ($\rho_{uz}=0$).

100 draws of size of 10000 are used. For each sample, $\hat{\rho}_{xu}$ *is varied in* $D=(0.15,0.5)$

3.3 Robustness Checks

To ensure methodological rigor, we implement alternative confidence interval estimation techniques for ρ_{uz}. Table 3 presents results using both the Bias-Corrected and Accelerated (BCa) and the Bootstrap Percentile Interval (BPI) methods. These complementary approaches corroborate our primary findings: the confidence intervals for z_p consistently include zero while those for z_n do not, providing strong evidence for the validity of z_p and the invalidity of z_n as instrumental variables.

Table 3: Alternative 95% Confidence Intervals for $\hat{\rho}_{uz}$ using Instruments z_p and z_n

Instrument	**95% BCa CI**	**95% BPI CI**
z_p	[-0.295, 0.074]	[-0.205, 0.189]
z_n	[0.048, 0.377]	[0.128, 0.481]

Notes: Results based on 100 bootstrap samples of size 10,000. For each sample, $\hat{\rho}_{xu}$ *is varied across the domain* $D=(0.15,0.5)$. *BCa (Bias-Corrected and Accelerated) and BPI (Bootstrap Percentile Interval) methods implemented using the R package "boot" (version 1.3-31).*

The consistency across different confidence interval estimation techniques strengthens our conclusions and demonstrates the CR test's robustness to methodological variations. Notably, the BPI intervals display slightly different bounds compared to the BCa intervals, yet they yield identical sub-

stantive conclusions regarding instrument validity. This invariance to the specific confidence interval methodology provides additional assurance of the test's reliability. The results further reinforce our earlier finding that the CR test effectively discriminates between valid and invalid instruments, with direct implications for the consistency of the resulting IV estimates.

3.4 Applying the IV correlation restrictions test to empirical data

We demonstrate the practical value of the CR test by examining Card's (1993) influential study on returns to education. Card estimated the following wage equation:

$$lwage_i = \beta \cdot educ_i + H_i' \cdot \alpha + u_i \tag{8}$$

where H_i' includes experience, race, location controls, and regional dummies. Card employed four instrumental variables: proximity to two-year (*nearc*2) and four-year colleges (*nearc*4), and parental education levels ($fatheduc$, *motheduc*). The endogeneity in this context is believed to stem primarily from omitted ability bias, which creates a positive correlation between education and the error term ($corr(educ, u) > 0$). Additional sources of endogeneity include potential misspecification and nonlinearity, particularly "sheepskin" effects associated with degree completion. This positive correlation suggests that OLS estimates should overstate the true returns to education.

The analysis reveals an intriguing pattern in the IV estimates. Using *nearc*4 as an instrument yields education returns (0.132) approximately 76% higher than the OLS estimate (0.075), while *nearc*2 produces an estimate more than three times larger (0.293), albeit statistically insignificant. The parental education instruments ($fatheduc$ and *motheduc*) generate statistically significant estimates 20-37% higher than OLS. These counter-intuitive results - where IV estimates exceed OLS despite positive ability bias - have prompted several explanations in the literature: measurement error creating downward OLS bias, heterogeneous returns to education affecting the local average treatment effect, invalidity of family background variables (Card, 1999). Notably, a recent study by Chalak (2019) builds on the work of Card (1995) and utilizes sign and magnitude restrictions on confounding to document a nonlinear return to education that is smaller than the ordinary least squares estimates.

To implement the CR test, we compute correlations using reduced form variables: $y = (I - P_H)lwage$, $x = (I - P_H)educ$, and $z = (I - P_H)\tilde{z}$ for each instrument $\tilde{z}$, where P_H is the linear projection matrix onto

Table 4: The correlation restrictions: Card example

	Parameter estimates		*Inference*		
IVs $\tilde{z}$	OLS: $\hat{\beta}$	IV: $\hat{\beta}_{IV}$	$P(\hat{\rho}_{uz}=0)>0$	95% CI:$\hat{\rho}_{uz}$	95% CI: $\hat{\beta}_{IV}$
	0.075^{***} (0.003)				
$fatheduc$		0.09^{***} (0.014)	0.24	[-0.033, 0.039]	[0.061, 0.118]
motheduc		0.103^{***} (0.014)	0.44	[-0.035, 0.034]	[0.076, 0.131]
*nearc*2		0.293 (0.185)	0.0	[0.004, 0.077]	[-0.082, 0.664]
*nearc*4		0.132^{**} (0.055)	0.0	[0.0097, 0.082]	[0.020, 0.242]

$P(\hat{\rho}_{uz}=0)>0$ gives the fraction of draws compatible with a valid instrument ($\rho_{uz}=0$).

50 draws of size of 2860 are used. For each sample, $\hat{\rho}_{xu}$ is varied in $D=(0,0.9)$

H. The results in Table 4 show that $fatheduc$ and *motheduc* pass the CR test, while *nearc*2 and *nearc*4 fail. However, the relatively low coverage probabilities for $fatheduc$ and *motheduc* (0.24 and 0.44 respectively) suggest these instruments are only weakly reliable. This weakness, combined with potential nonlinear effects highlighted by Chalak (2019), may explain their upward-biased estimates despite passing the CR test.

Table 5: 95% CI: $\hat{\rho}_{uz}$: BCa and BPI method

IV	**95% BCa CI: $\hat{\rho}_{uz}$**	**95% BPI CI: $\hat{\rho}_{uz}$**
fatheduc	[-0.062, 0.020]	[-0.041, 0.046]
motheduc	[-0.051, 0.011]	[-0.036, 0.031]
nearc2	[0.024, 0.044]	[0.029, 0.051]
nearc4	[0.005, 0.057]	[0.017, 0.073]

To verify the robustness of our findings, we employ both BCa and BPI confidence interval estimation methods for ρ_{uz} (Table 5). Both approaches corroborate our initial conclusions: confidence intervals for $fatheduc$ and *motheduc* include zero, while those for *nearc*2 and *nearc*4 do not. These results underscore a critical methodological insight: relying solely on theoretical arguments for instrument validity can lead to biased causal estimates. The CR test provides a valuable empirical diagnostic

when the sign of $corr(x,u)$ is known, helping researchers identify potentially invalid instruments and improve the reliability of their causal inference.

4 Applications Across Diverse Research Settings

To evaluate the generalizability and effectiveness of our Correlation Restriction (CR) test, we examine four influential empirical studies that employ instrumental variables across fundamentally different research contexts. Each application offers unique insights into the test's performance under varying endogeneity scenarios and data structures.

4.1 Electronic Monitoring and Recidivism

Di Tella and Schargrodsky (2013) investigate the causal impact of electronic monitoring (EM) versus imprisonment on criminal recidivism using the model:

$$R_i = \alpha + \beta \cdot EM_i + X_i'\gamma + \varepsilon_i \tag{9}$$

where R_i indicates post-release detention, EM_i denotes electronic monitoring assignment, and X_i incorporates controls for crime type, demographic characteristics, and judicial factors. Since judges likely assign EM to lower-risk offenders, there exists a plausible positive selection bias ($corr(EM,u) > 0$).

The authors employ two instruments: (1) the judge's EM assignment rate (excluding the particular offender) and (2) a binary indicator for whether the judge had previously used EM. Our CR test results indicate that both instruments satisfy validity conditions, though with moderate coverage probabilities (0.36 and 0.58 respectively). The wider confidence intervals for $\hat{\beta}_{IV}$ suggest these instruments, while technically valid, may suffer from limited strength. Notably, our analysis reveals that instrument validity holds only within a specific endogeneity range $\hat{\rho}_{xu} \in (0, 0.6)$, which aligns with theoretical expectations in this context.

4.2 School Access and Educational Outcomes

Burde and Linden (2013) examine how school proximity affects academic achievement in rural Afghanistan

using the specification:

$$Test\_score_i = \alpha + \beta \cdot Enrollment_i + H_i'\gamma + \varepsilon_i \tag{10}$$

where H_i contains demographic covariates. The study addresses endogeneity arising from unobserved parental characteristics that simultaneously influence both school enrollment decisions and children's academic performance ($corr(Enrollment, u) > 0$).

Using village school presence as an instrument, our CR test reveals this instrument fails validity requirements across the entire domain $\hat{\rho}_{xu} \in (0, 0.9)$, with zero coverage probability and a 95% confidence interval for $\hat{\rho}_{uz}$ of [0.301, 0.434] decisively excluding zero. This finding suggests a direct pathway through which village school location may affect test scores beyond its influence on enrollment - potentially through community-level effects, teacher quality variations, or school infrastructure differences.

4.3 Colonial Institutions and Economic Development

Banerjee and Iyer (2005) investigate the long-term developmental consequences of British colonial land revenue systems in India. Their research examines how areas where landlords held proprietary land rights exhibited systematically lower agricultural investments, reduced productivity, and diminished investments in health and education infrastructure compared to regions where cultivators maintained these rights.

Based on historical patterns suggesting downward OLS bias ($corr(x, u) < 0$), they instrument the non-landlord proportion using British conquest timing (1820-1856) - a period when colonial administrators systematically implemented non-landlord systems irrespective of district characteristics. Our CR test provides robust support for this instrument's validity, with 0.98 coverage probability across the plausible endogeneity range $\hat{\rho}_{xu} \in (-0.7, -0.4)$. The narrow confidence interval for $\hat{\beta}_{IV}$ [0.48, 0.69] further reinforces the reliability of the causal estimates in this application.

Table 6: Correlation Restriction Test Results Across Empirical Applications

Endogenous Regressor	**Parameter Estimates**		**Inference**		
& Instrument	OLS: $\hat{\beta}$	IV: $\hat{\beta}_{IV}$	$P(\hat{\rho}_{uz}=0)>0$	95% CI: ρ_{uz}	95% CI: $\hat{\beta}_{IV}$
1. Di Tella and Schargrodsky (2013): $corr(x,u)>0$, $\hat{\rho}_{xu}\in(0,0.6)$					
ER: *ElectronicMonitoring*	-0.095^{***} (0.025)				
IV: *JudgeSentToEM*		-0.079^{***} (0.022)	0.36	[-0.001, 0.025]	[-25.8, 25.2]
IV: *JudgeEverUsedEM*		-0.133^{***} (0.025)	0.58	[-0.001, 0.025]	[-3.5, 2.5]
2. Burde and Linden (2013): $corr(x,u)>0$, $\hat{\rho}_{xu}\in(0,0.9)$					
ER: *Enrolled*	0.855^{***} (0.06)				
IV: *School_In_Village*		1.30^{***} (0.121)	0.00	[0.301, 0.434]	[1.06, 1.53]
3. Banerjee and Iyer (2005): $corr(x,u)<0$, $\hat{\rho}_{xu}\in(-0.7,-0.4)$					
ER: *p_nland*	0.089^{***} (0.013)				
IV: *British_Revenue_Control*		0.58^{***} (0.052)	0.98	[-0.036, 0.016]	[0.48, 0.69]
4. Galiani et al. (2011): $corr(x,u)<0$, $\hat{\rho}_{xu}\in(-0.03,0)$					
ER: *Conscription*	0.0031^{*} (0.0014)				
IV: *Draft_Eligible*		0.0043^{**} (0.0014)	0.98	[-0.032, 0.025]	[0.001, 0.007]

Notes: $P(\hat{\rho}_{uz}=0)>0$ represents the fraction of bootstrap draws compatible with a valid instrument ($\hat{\rho}_{uz}=0$). ER denotes the endogenous regressor. Standard errors in parentheses. Significance: $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.1$.

4.4 Military Service and Criminal Behavior

Galiani et al. (2011) estimate the effect of mandatory military service on subsequent criminal behavior:

$$CrimeRate_{ci}=\alpha+\beta\cdot Conscription_{ci}+\delta_c+\varepsilon_{ci} \tag{11}$$

where δ_c represents cohort fixed effects. They employ draft lottery eligibility as an instrument to address negative selection bias ($corr(x,u)<0$) arising from systematic draft avoidance behaviors.

While our CR test strongly supports this instrument's validity with 0.98 coverage probability, this conclusion remains sensitive to the presumed magnitude of endogeneity. Specifically, validity holds only within the narrow range $\hat{\rho}_{xu}\in(-0.03,0)$, indicating that even modest deviations in endogeneity

strength could potentially compromise the instrument's exogeneity. This finding illustrates the value of the CR test in identifying critical sensitivity thresholds for instrumental variables.

Table 7: Alternative 95% Confidence Intervals for $\hat{\rho}_{uz}$ Across Applications

Instrument	**BCa Method**	**BPI Method**
1. Di Tella and Schargrodsky (2013): $corr(x,u) > 0$, $\hat{\rho}_{xu} \in (0, 0.6)$		
Judge Sent to EM	[-0.091, 0.047]	[-0.057, 0.091]
Judge Ever Used EM	[-0.079, 0.039]	[-0.050, 0.077]
2. Burde and Linden (2013): $corr(x,u) > 0$, $\hat{\rho}_{xu} \in (0, 0.9)$		
School in Village	[0.039, 0.448]	[0.139, 0.575]
3. Banerjee and Iyer (2005): $corr(x,u) < 0$, $\hat{\rho}_{xu} \in (-0.7, -0.4)$		
British Revenue Control	[-0.115, 0.017]	[-0.083, 0.059]
4. Galiani et al. (2011): $corr(x,u) < 0$, $\hat{\rho}_{xu} \in (-0.03, 0)$		
Draft Eligible	[-0.032, 0.004]	[-0.023, 0.015]

Notes: BCa (Bias-Corrected and Accelerated) and BPI (Bootstrap Percentile Interval) confidence intervals for instrument correlation with the structural error term. Consistent conclusions across both methods support the robustness of the CR test findings.

In summary, our cross-study analysis demonstrates the practical value of the CR test in empirical research settings. The test not only identifies potentially invalid instruments (as in the Burde and Linden study) but also reveals the critical range of endogeneity magnitudes across which instruments maintain their validity (notably in the Galiani et al. application). The consistency of our findings across different bootstrap confidence interval methods, as confirmed by both BCa and BPI approaches (Table 7), further substantiates the robustness of the CR test as a diagnostic tool for instrumental variable validity when the direction of endogeneity is known.

5 Conclusion

This paper develops and validates a novel Correlation Restriction (CR) test for evaluating instrumental variable validity when the direction of endogeneity is known. By leveraging the theoretical constraint

that any instrument must satisfy, we provide applied researchers with a practical empirical tool to assess instrument exogeneity under partial identification conditions.

Our methodological contribution builds directly on the theoretical framework established by Di-Traglia and García-Jimeno (2021), who demonstrate that when the sign of $corr(x,u)$ is known, valid instruments must adhere to a specific correlation restriction with observable variables. Any violation of this restriction definitively indicates instrument invalidity, as it necessarily implies $E(z'u) \neq 0$. Through simulations with controlled data-generating processes, we establish that instruments failing the CR test systematically produce biased parameter estimates, validating the test's discriminatory power in finite samples.

The empirical applications reveal the substantive value of our approach in real-world research settings. Revisiting Card (1993)'s seminal study on returns to education, we find that two instruments - proximity to two-year and four-year colleges – fail the CR test, explaining their contribution to estimation bias. The remaining instruments - parental education measures – while technically passing the test, exhibit relatively low coverage probabilities (0.24 and 0.44), indicating limited reliability. These findings align with recent literature questioning traditional specifications in returns to education models and highlight the importance of rigorous instrument validation.

Our analysis across four additional empirical applications - spanning criminal justice, development economics, education, and labor economics - further demonstrates the test's broad utility. The CR test successfully validates instruments in three cases while identifying invalidity in the fourth. Notably, even among valid instruments, the test reveals meaningful heterogeneity in reliability through coverage probabilities and precisely delineates the range of endogeneity magnitudes over which instruments maintain validity. This nuanced assessment provides researchers with critical information beyond binary validity judgments.

These findings carry significant implications for empirical practice. When researchers can determine the direction of endogeneity through economic theory, institutional knowledge, or contextual expertise, the CR test offers a valuable diagnostic tool for instrument selection and validation. The test not only helps researchers avoid invalid instruments but also provides a quantitative assessment of instrument reliability within specific endogeneity ranges. This demonstrates the essential comple-

mentarity between substantive economic theory and formal statistical testing in credible instrumental variables research.

Future research could productively extend this framework in several directions. First, developing systematic methods for empirically determining the sign of $corr(x,u)$ would substantially expand the test's applicability across diverse research contexts. Second, extending the framework to accommodate partial identification of the endogeneity sign could enhance its practical value in settings where theoretical predictions are ambiguous. Finally, integrating the CR test with other instrument validity diagnostics, such as tests for instrument relevance and monotonicity, could provide a more comprehensive assessment framework for causal inference with instrumental variables.

Funding details.

The authors report that there is no funding to be declared.

Disclosure statement

The authors report there are no competing interests to declare.

Data availability statement

1. Card (1993), Using Geographic Variation in College Proximity to Estimate the Return to Schooling. NBER Working Papers 4483, National Bureau of Economic Research, Inc. The dataset is available from "http://qcpages.qc.cuny.edu/ rvesselinov/statadata/CARD.DTA".

2. Burde and Linden (2013). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. American Economic Journal: Applied Economics 5, 27–40. The dataset is available from the journal website: http://doi.org/10.3886/E113861V1

3.Galiani et al. (2011). Conscription and Crime: Evidence from the Argentine Draft Lottery. American Economic Journal: Applied Economics 3(2), 119–136. The dataset is available from the journal website: http://doi.org/10.3886/E113781V1

4. Banergee and Iyer (2005). History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India. American Economic Review, 95 (4): 1190-1213. The dataset is available from the journal website: http://doi.org/10.3886/E116059V1

5. Di Tella, R., E. Schargrodsky (2013). Criminal Recidivism after Prison and Electronic Monitoring. Journal of Political Economy,121(1), 28–73.The dataset is available from the journal website: https://doi.org/10.1086/669786

Appendix

Extension to Overidentification and Multiple Endogenous Regressors

While the main analysis focuses on regression models with a single endogenous regressor in the exactly identified case, this appendix extends our framework to more general settings with overidentification and multiple endogenous regressors.

Overidentified Models with a Single Endogenous Regressor

Consider an extension of the model specified in equations (4) and (5):

$$y = \beta x + v, \tag{12}$$

$$x = \pi_1 z_1 + \pi_2 z_2 + \cdots + \pi_n z_n + \varepsilon, \tag{13}$$

where we maintain a single endogenous regressor x but now incorporate n instrumental variables: $z_1, z_2, \ldots, z_n$.

To evaluate instrument validity in this overidentified setting, we must account for the potential interdependence among instruments. Our approach requires that each instrument individually satisfies the Correlation Restriction (CR) test. For each instrument z_j, we apply the CR test methodology developed in the main text, examining whether the correlation between z_j and the structural error term u is consistent with zero given the assumed sign and magnitude of endogeneity $corr(x,u)$.

Models with Multiple Endogenous Regressors

We now extend the framework to accommodate models with multiple endogenous regressors where the number of instruments exceeds the number of endogenous variables:

$$\tilde{y} = \tilde{X}\beta + H'\gamma + v, \tag{14}$$

$$\tilde{X} = Z\Pi + \varepsilon, \tag{15}$$

where $\tilde{X}$ is an $n \times k$ matrix of endogenous regressors, H is a matrix of exogenous variables including a constant term, and $Z = [\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n]$ is a matrix of n instrumental variables with $n > k$.

The key insight for extending our methodology to this setting is to decompose the multiple endogenous variable case into a series of single endogenous variable problems through appropriate projections. We proceed as follows:

1. **Instrument Assignment**: For each endogenous variable $\tilde{x}_i$ ($i = 1, 2, \ldots, k$), we assign a specific subset of instruments $Z_i \subset Z$. This partitioning ensures that each endogenous regressor has its own dedicated set of instruments, facilitating the identification of individual causal effects.

2. **First-Stage Projections**: For each endogenous variable $\tilde{x}_i$, we compute its projection onto its corresponding instrument space: $\hat{x}_i = P_{Z_i}\tilde{x}_i$, where P_{Z_i} denotes the linear projection matrix onto the space spanned by Z_i. This yields the matrix of fitted values $\hat{X}$.

3. **Sequential Residualization**: For each endogenous variable $\tilde{x}_i$, we define $H_i = [H|\hat{X}_{-\hat{x}_i}]$, where $\hat{X}_{-\hat{x}_i}$ represents the matrix $\hat{X}$ excluding the column $\hat{x}_i$. We then compute residuals by projecting the relevant variables onto the orthogonal complement of the space spanned by H_i:

$$z_j = (I - P_{H_i})\tilde{z}_j \quad \text{for each } z_j \in Z_i, \tag{16}$$

$$x_i = (I - P_{H_i})\tilde{x}_i, \tag{17}$$

$$y = (I - P_{H_i})\tilde{y}. \tag{18}$$

4. **Application of CR Test**: After this residualization, the relationship between the transformed

variables can be expressed as:

$$y = \beta_i x_i + v_i, \tag{19}$$

$$x_i = \sum_{z_j \in Z_i} \pi_j z_j + \varepsilon_i. \tag{20}$$

This transformation effectively reduces the multiple endogenous regressor case to a series of single endogenous regressor models, allowing us to apply the CR test to each instrument $z_j \in Z_i$ individually.

This procedure is then repeated for each endogenous regressor $\tilde{x}_i$, enabling comprehensive validity assessment for all instruments across all endogenous variables. This approach provides a systematic framework for applying the CR test in more complex models while maintaining the interpretability and theoretical foundations established in the main text.

References

Acerenza, S., O. Bartalotti, and D. Kedagni (2023). Testing identifying assumptions in bivariate probit models. *Journal of Applied Econometrics 38*(3), 407–422. `https://doi.org/10.1002/jae.2956`.

Banerjee, A., and L. Iyer. (2005). History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India. *American Economic Review 95*(4), 1190–1213.

Berkowitz, D., M. Caner, and Y. Fang (2012, feb). The validity of instruments revisited. *Journal of Econometrics 166*(2), 255–266.

Beasley, W. H., L. DeShea, L. Toothajer, J. Mendoza, D. Bard, and J. Rodgers (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological Methods 12*(4), 414–433. doi: 10.1037/1082-989X.12.4.414

Bonet, B. (2001). Instrumentality tests revisited. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*. Burlington, MA: Morgan Kaufmann.

Burde, D., and L. Linden (2013). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics 5*, 27–40.

Card, D. (1993, October). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. NBER Working Papers 4483, National Bureau of Economic Research, Inc.

Card, D. (1999). *The Causal Effect of Education on Earnings* Chapter 30 in the Handbook of Labor Economics. (Volume 3A). Amsterdam: Elsevier.

Chan, W. and D. W.-L. Chan (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods 9*(3), 369–385. doi: 10.1037/1082-989X.9.3.369

Chen, X., T. M. Christensen, and E. Tamer (2018). Monte Carlo confidence sets for identified sets. *Econometrica 86*, 1965–2018. https://doi.org/10.3982/ECTA14525.

Chalak, K. (2019). Identification of average effects under magnitude and sign restrictions on confounding. *Quantitative Economics 10*, 1619–1657. https://doi.org/10.3982/QE689

Davison, A. C. and D. V. Hinkley(1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

D'Haultfœuille, X., S. Hoderlein and Y. Sasaki (2021). Testing and relaxing the exclusion restriction in the control function approach. *Journal of Econometrics* forthcoming. https://www.sciencedirect.com/science/article/pii/S0304407621000439

Di Tella, R., E. Schargrodsky (2013). Criminal Recidivism after Prison and Electronic Monitoring. *Journal of Political Economy 121*(1), 28–73.

DiTraglia, F. J. and C. García-Jimeno (2021). A framework for eliciting, incorporating, and disciplining identification beliefs in linear models. *Journal of Business & Economic Statistics 39*(4), 1038–1053.

Efron, B. and R.J. Tibshirani (1994). *An Introduction to the Bootstrap* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Fisher, R. A. (1921). Probable Error of a Coefficient of Correlation Deduced from a Small Sample. *Metron 1*, 3–32.

Fujikoshi, Y., V. V. Ulyanov, and R. Shimizu (2010, jan). *Multivariate Statistics*. John Wiley & Sons, Inc., New Jersey, USA.

Galiani, S., M.A. Rossi, and E. Schargrodsky. (2011). Conscription and Crime: Evidence from the Argentine Draft Lottery. *American Economic Journal: Applied Economics 3*(2), 119–136.

Gunsilius, F. F. (2021). Nontestability of instrument validity under continuous treatments. *Biometrika 108*(4), 989–995.

Howell, D.C. (2003). *Statistical Methods for Psychology* (Fifth ed.). Thompson Wadsworth, Belmont, USA.

Jiang, Z. and P. Ding (2020). Measurement errors in the binary instrumental variable model. *Biometrika 107*, 238–245.

Kédagni, D. and I. Mourifié (2020). Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika 107*, 661–675.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*, 2043–2063.

Kitagawa, T. (2012). Estimation and inference for set-identified parameters using posterior lower probability. Working Paper, available at http://www.homepages.ucl.ac.uk/~uctptk0/Research/LowerUpper.pdf.

Kiviet, Jan F. (2020). Testing the impossible: Identifying exclusion restrictions. *Journal of Econometrics 218*(2), 294–316.

Kline, B. and E. Tamer (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics 7*, 329–366.

Lehmann, E. L. (1998). *Elements of large sample theory*. New York: Springer.

Li, J. C.-H. (2022). Bootstrap confidence intervals for 11 robust correlations in the presence of outliers and leverage observations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 18*(2), 99–125. doi: 10.5964/meth.8467

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer.

Manski, C. F. and J. V. Pepper (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica 68*(4), 997–1010.

Moon, H. R., and F. Schorfheide (2009). Estimation With Overidentifying Inequality Moment Conditions. *Journal of Econometrics 153*, 136–154.

Nevo, A. and A. M. Rosen (2012, aug). Identification with imperfect instruments. *Review of Economics and Statistics 94*(3), 659–671.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*, 669–688.

Puth, M.-T., M. Neuh auser, and G. D. Ruxton (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology 84*, 892–897. doi: 10.1111/1365-2656.12382

Silva, R. and S. Shimizu (2017). Learning instrumental variables with structural and non-Gaussianity assumptions. *Journal of Machine Learning Research 18*, 1–49.

Xie, F., Y. He, Z. Geng, Z. Chen, R. Hou, and K. Zhang (2022). Testability of instrumental variables in linear non-Gaussian acyclic causal models. *Entropy (Basel) 24*(4), 512. `https://doi.org/10.3390/e24040512`.

Xie, F., R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang (2020). Generalized independent noise condition for estimating latent variable causal graphs. In *Proceedings of Advances in Neural Information Processing Systems*, Virtual, December 6–12, pp. 14891–14902.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Wang, L., J. M. Robins, and T. S. Richardson (2017). On falsification of the binary instrumental variable model. *Biometrika 104*, 229–236.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5 ed.). South-Western Cengage Learning.