An Improved Inference for IV Regressions*

Liyu Dou[a], Pengjin Min[a], Wenjie Wang[b], and Yichong Zhang[a]

[a] *Singapore Management University, Singapore*

[b] *Nanyang Technological University, Singapore*

Abstract

Empirical instrumental variables (IV) studies often report separate results based on low-dimensional instruments and many base instruments. This paper proposes a combination test that integrates these commonly reported statistics. The test linearly combines a cluster-robust Wald statistic based on low-dimensional IVs with leave-one-cluster-out Lagrangian multiplier (LM) and Anderson-Rubin (AR) statistics constructed from many IVs. We establish joint asymptotic normality and asymptotic optimality of the proposed test. The procedure yields costless efficiency improvements, automatically adapts to weak identification of many instruments, and is accompanied by a practical rule of thumb for assessing efficiency gains.

Keywords: Many Weak Instruments, Shift-Share Instruments, Combination Test.

JEL Classification: C12, C36, C55

*This version is dated on March 10, 2026. We are grateful to Michal Kolésar for helpful comments. Dou acknowledges financial support from the National Natural Science Foundation of China through Grants 72133005 and 72573149. Wang acknowledges the support from the Singapore Ministry of Education Tier 1 Grant RG51/24. Zhang acknowledges support from the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (Project ID: MOE-000767-00). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore. Any and all errors are our own.

1 Introduction

Empirical applications of instrumental variables (IV) regressions in economics often involve multiple sets of candidate instruments, some with dimensionality proportional to the sample size and others more parsimonious. A canonical example is the influential study by Angrist and Krueger (1991), which estimates the causal effect of schooling on wages using three quarter-of-birth (QoB) dummies as instruments, as well as their interactions with state- and year-of-birth dummies, yielding a total of 1,530 instruments. A different but related class of examples arises in the context of shift-share IVs,[1] which, as noted by Goldsmith-Pinkham et al. (2020), can be interpreted as a particular way of aggregating many underlying base instruments. It is common practice to report results based on a one-dimensional shift-share IV and then supplement them with results obtained using the full set of base IVs.

This naturally raises the question of whether one can take the low-dimensional IV regression as the benchmark specification and combine it with its many-IV counterpart in a way that enhances the power of IV inference, for free, in the sense that no additional identification restrictions are imposed on the many-IV specification. In this paper, we provide a constructive solution.

Specifically, we consider an optimal combined inference procedure based on three commonly reported test statistics for IV regressions in a clustered setting, where the data consist of many clusters of bounded size. The three component statistics are: (i) a standard cluster-robust Wald statistic constructed from the low-dimensional IVs; (ii) a leave-one-cluster-out jackknife Lagrange multiplier (LM) statistic; and (iii) a leave-one-cluster-out jackknife Anderson–Rubin (AR) statistic. The LM and AR statistics are constructed using the many-IV specification, and the leave-one-cluster-out design removes the many-IV bias in the presence of within-cluster error dependence.

[1] See, for instance, Bartik (1991), Blanchard, Katz, Hall, and Eichengreen (1992), Adao, Kolesár, and Morales (2019), Goldsmith-Pinkham, Sorkin, and Swift (2020), Borusyak, Hull, and Jaravel (2022), Borusyak, Hull, and Jaravel (2025), and references therein.

We first show that, under the null hypothesis and local alternatives, the Wald, LM, and AR statistics are jointly asymptotically normal, provided that the low-dimensional IVs strongly identify the parameter of interest, while allowing the many-IV specification to be weakly identified. Standard optimal testing theory then implies that, in the corresponding Gaussian limit experiment, the uniformly most powerful unbiased (UMPU) test rejects for large absolute values of an appropriate linear combination of the three limiting Gaussian statistics.[2] Our proposed test, as a function of the three statistics, is asymptotically equivalent to this UMPU test and is therefore weakly more powerful than each of the Wald, LM, and AR tests.

Importantly, the combination test adapts automatically to the identification strength of the many-IV specification. When the parameter of interest is weakly identified under many IVs in the sense of Mikusheva and Sun (2022), the optimal linear combination places asymptotically negligible weight on the LM and AR components, so that the resulting test reduces to the Wald test. In this case, the procedure remains valid and retains the same asymptotic power as the standalone Wald test.

The notion of optimality in our study is defined relative to the class of tests based on the Wald, LM, and AR statistics and serves primarily as a device for improving inference within this class. We do not attempt to derive a globally optimal test by searching over all possible combinations of a given set of base IVs, nor do we aggregate results across alternative specifications that employ different sets of base IVs or different low-dimensional IVs. Rather, our objective is to take the low-dimensional specification as the benchmark and combine its associated test statistic with those constructed from a given set of many IVs, thereby strengthening inference while remaining agnostic about the identification strength of the many-IV specification. Nonetheless, we attribute a notion of asymptotic optimality to the resulting combination test in the sense of Müller (2011).

[2] See, for instance, Section 4.2 of Lehmann and Romano (2006) and Lemma 2.2 of Lim, Wang, and Zhang (2024a) for formal arguments.

The confidence interval implied by the combination test has the usual "estimator plus and minus a standard error times a critical value" form. Its center is an estimator that linearly combines a standard GMM estimator based on the low-dimensional IVs with a leave-one-cluster-out jackknife IV estimator based on many IVs and the AR statistic, using weights that capture both the relative identification strength from the two IV sets and the UMPU weights. The efficiency gain manifests itself as an almost surely shorter confidence interval. We measure this gain by the percentage reduction in the length of the resulting confidence interval relative to that of the Wald test based on the low-dimensional IVs. As an illustration, in the immigrant enclave application of Card (2009), the combination procedure reduces the length by between 5% and 17%. This reduction depends mainly on the identification strengths of the low-dimensional and many IVs, together with the limiting correlations between the Wald and LM statistics, and between the LM and AR statistics. In Section 2, we translate these relationships into a practical rule of thumb based solely on the ratio of the standard errors of the low-dimensional and many-IV estimators, which can be directly read from reported regression tables prior to implementing our combination procedure. We illustrate this rule using the application in Card (2009).

Relation to the literature. This paper contributes to the large literature on many (weak) instruments,[3] and is particularly related to Lim et al. (2024a). Building on Andrews (2016), they propose a jackknife conditional linear combination (CLC) test for independent data

[3]See, for instance, Kunitomo (1980), Morimune (1983), Bekker (1994), Donald and Newey (2001), Chao and Swanson (2005), Stock and Yogo (2005), Han and Phillips (2006), Andrews and Stock (2007), Hansen, Hausman, and Newey (2008), Newey and Windmeijer (2009), Anderson, Kunitomo, and Matsushita (2010), Kuersteiner and Okui (2010), Anatolyev and Gospodinov (2011), Okui (2011), Belloni, Chen, Chernozhukov, and Hansen (2012), Carrasco (2012), Chao, Swanson, Hausman, Newey, and Woutersen (2012), Hausman, Newey, Woutersen, Chao, and Swanson (2012), Kolesár (2013), Hansen and Kozbur (2014), Carrasco and Tchuente (2015), Wang and Kaffo (2016), Kolesár (2018), Evdokimov and Kolesár (2018), Sølvsten (2020), Chao, Swanson, and Woutersen (2023), Lim et al. (2024a), Boot and Nibbering (2024), Yap (2025), Lim, Wang, and Zhang (2024b), among others.

that is robust to weak identification, many instruments, and heteroskedasticity, combining jackknife AR, LM, and orthogonalized LM statistics to achieve good power across identification regimes. Their analysis, however, is confined to many-IV settings and does not directly apply to our framework. In contrast, we explicitly incorporate low-dimensional IVs within a clustered environment and develop a unified procedure that efficiently combines estimation and inference from both low- and high-dimensional IV specifications. Our primary objective is to construct a theoretically grounded combination test that improves upon the conventional Wald test based solely on low-dimensional IVs. We also note that, in a different context, Jiang, Li, Miao, and Zhang (2025) proposes an optimal linear combination of adjusted and unadjusted estimators for the average treatment effect under covariate-adaptive randomization.

Our paper is also related to the literature on inference with many IVs and clustered data. Frandsen, Leslie, and McIntyre (2025) study leave-one-cluster-out jackknife versions of the IV estimator, and Ligtenberg (2025) develop weak-identification-robust procedures in clustered environments. We depart from this literature by not focusing solely on estimation or inference within the many-IV specification itself. Instead, we leverage the many-IV specification to enhance the power of low-dimensional Wald inference through an optimal combination of test statistics. Chao et al. (2023) and Kolesár, Min, Wang, and Zhang (2026) consider settings with both many IVs and many control variables, which introduce additional challenges for cluster-robust inference. In this paper, we restrict attention to models with a fixed number of control variables and leave the case with many IVs and many controls for future research.

Structure of the paper: Section 2 details the rationale behind the proposed rule of thumb and demonstrates its use in an empirical application; practitioners mainly interested in applications may focus on this section directly. Section 3 introduces the model and key preliminaries, while Section 4 develops the large-sample theory for the combination test and formalizes its theoretical properties. Section 5 presents the practical implementation of the combination test in an empirically relevant setting and employs simulations to evaluate its power properties, and Section 6 concludes. An additional case with weak low-dimensional IVs alongside strong many IVs, as well as all proofs, is presented in the Online Appendices.

Notation. We write $[n] \equiv \{1, \ldots, n\}$ and $[G] \equiv \{1, \ldots, G\}$. Let A be an $n \times m$ matrix and let $\{n_g\}_{g \in [G]}$ be positive integers with $\sum_{g=1}^{G} n_g = n$. We denote by $A_{[g]}$ the g-th row-wise block of A, of dimension $n_g \times m$. When A is an $n \times n$ square matrix, we denote by $A_{[g,h]}$ the (g, h)-th block of A. For a positive semi-definite square matrix A, denote its largest and smallest eigenvalues by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. Let C be a generic positive constant independent of n, whose value may change from line to line. For brevity, we write $\sum_{g,h \in [G]^2, g \neq h} := \sum_{g \in [G]} \sum_{h \in [G], h \neq g}$.

2 Rule of Thumb and Empirical Illustration

In this section, we develop a practical rule of thumb that can be directly applied to reported estimates and standard errors from regressions using low-dimensional IVs and from regressions employing many IVs. We illustrate its empirical relevance using the application in Goldsmith-Pinkham et al. (2020, Section VII), which builds on Card (2009).

We quantify the efficiency gain from the combination test by the percentage reduction in the length of its confidence interval, relative to that of the Wald test, in large samples. Section 4.3 formally derives an analytical expression for this reduction as a function of the identification strengths of both the low-dimensional and many instrumental variables (IVs), as well as the limiting correlations between the Wald and LM statistics and between the LM and AR statistics. Because the percentage reduction is monotonically increasing in the absolute correlation between the LM and AR statistics, we further obtain a lower bound on this reduction by setting this correlation to zero. This yields a lower bound on the efficiency gain that depends only on the correlation between the Wald and LM statistics (denoted as ρ_1) and on the ratio of the standard deviations of the GMM estimator based on low-dimensional IVs to the leave-one-cluster-out jackknife IV estimator (JIVE) based on many

IVs (or, equivalently, on the relative strength of the many IVs to the low-dimensional IVs).

Figure 1 plots the lower bound as a function of the standard deviation ratio for different values of ρ_1, the limiting correlation between the Wald and LM statistics. Two observations emerge. First, for any fixed ρ_1, we show theoretically that whenever the standard deviation ratio exceeds ρ_1, the lower bound on efficiency gains increases with the ratio, reflecting the fact that the combination test exploits the additional precision provided by the LM statistic. Second, once the standard deviation ratio exceeds one, the lower bound decreases as ρ_1 increases, because the LM statistic then contributes relatively little additional information beyond the highly correlated Wald statistic. As a simple rule of thumb that only requires a back of envelope calculation based on the reported standard errors, we propose: for empirically plausible values of ρ_1 (between -0.7 and 0.7), whenever the standard error from the regression with low-dimensional IVs divided by that from the regression with many IVs is greater than 1.05, the corresponding confidence interval is reduced by at least 10%.[4] We further argue in Section 4.3 that the same rule-of-thumb can be applied to other many-IV estimators such as HLIML and HFUL developed by Hausman et al. (2012).

To illustrate the empirical relevance of efficiency gains and the associated rule of thumb, we implement the combination test in an empirical application that estimates the (negative) inverse elasticity of substitution between immigrants and natives, following Card (2009). As in Goldsmith-Pinkham et al. (2020, Section VII), we examine two separate sets of results by skill group: high school equivalent workers and college equivalent workers. The analysis is based on cross-sectional regressions for each skill group in 124 cities in 2000. The dependent variable is the residual log wage gap between immigrant and native men, and the main regressor of interest is the log ratio of immigrant-to-native hours for both men and women within the same skill group. Because a positive labor demand shock to immigrants can simultaneously increase their earnings and labor supply relative to natives, this introduces

[4]In additional (unreported) plots for $\rho_1 \in [-0.99, 0.99]$, the confidence interval is at least 10% shorter whenever the standard deviation ratio exceeds 1.11 (with thresholds 1.05 for 5% and 1.25 for 20%).

potential endogeneity. To construct the one-dimensional Bartik instrument (i.e., shift-share instrument), immigration shares from 38 origin countries (groups) in 1980 are used as the base instruments, and the final instrument is formed as a weighted average of these country-specific shares, where the weights are given by the number of arrivals to the United States between 1990 and 2000 by origin country group and skill group (see Goldsmith-Pinkham et al. (2020) for further details). As argued in Card (2009), the rationale for the IV is that existing immigrant enclaves are likely to attract additional immigrant labor through social and cultural channels unrelated to labor market outcomes.

Table 1 reports the point estimates obtained from the two-stage least squares (TSLS) estimator using the Bartik instrument, $\hat{\beta}_1$, along with those from the leave-one-cluster-out estimator, $\hat{\beta}_2$, which relies on all the 38 base IVs. We present results separately for specifications that include and exclude city-level controls. As expected and consistent with the findings in Goldsmith-Pinkham et al. (2020), the estimates are broadly similar within each skill group. However, in every specification, the confidence intervals constructed from $\hat{\beta}_1$ (Wald CI) differ to some extent from those based on $\hat{\beta}_2$ (LM CI). This discrepancy is partly due to the different standard errors of the two estimators, which we exploit in our combination test to obtain strictly shorter confidence intervals across all specifications. For example, for workers with college equivalent skills, our confidence intervals are roughly 7% and 17% shorter, respectively, than the Wald CIs in the specifications with and without controls.

Figure 2 displays the realized percentage reduction in confidence interval length achieved by our combination test, together with the asymptotic lower bounds for that reduction, analogous to Figure 1, but calculated using the specification-specific estimate $\hat{\rho}_1$ of the limiting correlation between the Wald and LM statistics. Two observations are worth noting. First, up to finite-sample estimation error and across different specifications, the actual percentage reductions generally fall near or above the corresponding lower bounds, in line with our theoretical predictions. Note also that these bounds are derived solely from the Wald and LM statistics, indicating that, in this empirical setting, incorporating the AR statistic yields little additional efficiency gain. This aligns with our earlier theoretical discussion, which posits that efficiency gains increase monotonically with the absolute correlation between LM and AR statistics, and, in fact, the corresponding consistent estimates $\hat{\rho}_2$ are not particularly large in almost every specification.

Second, Figure 2 clearly illustrates our proposed rule of thumb. All estimated $\hat{\rho}_1$ values fall within $[-0.7, 0.7]$. When the standard error ratio exceeds 1.05, the actual reduction in the length of the confidence interval is 17% (specification without controls for college equivalent workers), much above the rule-of-thumb benchmark 10% for its lower bound. In contrast, when the standard error ratio remains at or below 1.05, the improvements are modest, reflecting the converse of our rule of thumb. Nevertheless, even in these cases, our combination test can still deliver confidence intervals that are about 7% (specification with controls for college equivalent workers), 6% (specification without controls for high school equivalent workers), and 5% (specification with controls for high school equivalent workers) shorter. In Supplementary Appendix F, we further consider the return to education application of Angrist and Krueger (1991), which illustrates even more notable efficiency gains provided by our approach.

3 Model and Preliminaries

3.1 Setup

We consider a clustered dataset with G clusters and denote the size of the g-th cluster as n_g for $g \in [G]$. We index observations by units followed by clusters. Denote $I_g = [N_{g-1}+1, \cdots, N_g]$, where $N_0 = 0$, $N_g = \sum_{g'=0}^{g} n_{g'}$, and $N_G = n$. Then, $\{I_g\}_{g\in[G]}$ forms a partition of $[n]$, and if $i \in I_g$, this means that the i-th observation belongs to the g-th cluster. We then consider a linear IV regression with clustered data:

$$\tilde{Y}_{i,g} = \tilde{X}_{i,g}\beta + W_{i,g}^{\top}\gamma + \tilde{e}_{i,g}, \tag{3.1}$$

where we denote $\tilde{Y}_{i,g} \in \Re$, $\tilde{X}_{i,g} \in \Re$, and $W_{i,g} \in \Re^{d_w}$ as an outcome variable, an endogenous regressor, and exogenous regressors, respectively. Further denote $\tilde{Z}_{i,g} \in \Re^K$ as the IVs for $\tilde{X}_{i,g}$. The first-stage equation can be written as

$$\tilde{X}_{i,g} = \tilde{\Pi}_{i,g} + \tilde{V}_{i,g}, \tag{3.2}$$

where $\tilde{\Pi}_{i,g} = \mathbb{E}(\tilde{X}_{i,g}|\{\tilde{Z}_{j,g}, W_{j,g}\}_{j\in I_g})$ is not assumed to be linear in $\tilde{Z}_{i,g}$ and $W_{i,g}$. We assume that $\mathbb{E}\tilde{e}_{i,g} = 0$ and $\mathbb{E}\tilde{V}_{i,g} = 0$, and $\left\{\tilde{e}_{i,g}, \tilde{V}_{i,g}\right\}_{i\in I_g, g\in[G]}$ are independent between clusters, but allow them to have a general dependence structure within each cluster. Throughout the paper, the dimension d_w of $W_{i,g}$ is assumed to be fixed. If researchers want to include cluster fixed effects in the model, they can obtain (3.1) by first demeaning the data (outcome, endogenous regressor, controls, and instruments) at the cluster level. We assume that K, the dimension of $\tilde{Z}_{i,g}$, diverges to infinity with the sample size.

Let $\tilde{Y}$, $\tilde{X}$, $\tilde{\Pi}$, W, $\tilde{Z}$ be $n \times 1$, $n \times 1$, $n \times 1$, $n \times d_w$, and $n \times K$-dimensional vectors and matrices formed by $\tilde{Y}_{i,g}$, $\tilde{X}_{i,g}$, $\tilde{\Pi}_{i,g}$, $W_{i,g}$, and $\tilde{Z}_{i,g}$, respectively. More specifically, $\tilde{Y}$ is constructed by stacking up $\tilde{Y}_{i,g}$ across $i \in I_g$ followed by $g \in [G]$, and similarly for $\tilde{X}$, $\tilde{\Pi}$, W and $\tilde{Z}$. We then partial out W from $\tilde{Y}$, $\tilde{X}$, and $\tilde{Z}$, so that the model in (3.1)-(3.2) can be written in a vector form as

$$Y = X\beta + e, \quad X = \Pi + V, \tag{3.3}$$

where $Y = M_W\tilde{Y}$, $X = M_W\tilde{X}$, $\Pi = M_W\tilde{\Pi}$, $e = M_W\tilde{e}$, $V = M_W\tilde{V}$, $M_W = I_n - P_W$, $P_W = W(W^\top W)^{-1}W^\top$, and I_n denotes an $n \times n$ identity matrix. We further denote $Z = M_W\tilde{Z}$.

In addition, besides the K-dimensional many IVs $\tilde{Z}_{i,g} \in \Re^K$, we assume that there is another set of low-dimensional IVs

$$\tilde{z}_{i,g} = f_{i,g}(\tilde{Z}, W) \in \Re^{d_z},$$

where $\{f_{i,g}(\cdot)\}_{i\in I_g, g\in[G]}$ is a list of known nonstochastic functions of d_z dimension. Specifically, as illustrated by the example of Angrist and Krueger (1991) in the Introduction, researchers may begin with certain low-dimensional base IVs $\tilde{z}_{i,g}$, such as the three QoB dummies, and construct a large number of new IVs by taking the interaction between $\tilde{z}_{i,g}$ and control variables $W_{i,g}$ (e.g., state- and year-of-birth dummies in Angrist and Krueger (1991)). Then, $\tilde{z}_{i,g}$ is a subset of the K-dimensional many IVs $\tilde{Z}_{i,g}$ for the model in (3.1)-(3.2), which include both the low-dimensional base IVs and interacted IVs. The second example of $\tilde{z}_{i,g}$ is the widely used shift-share IV. As pointed out by Goldsmith-Pinkham et al. (2020), under their identification strategy that treats the shares as exogenous, the one-dimensional shift-share IV can be regarded as a weighted average of many base IVs. For instance, in the canonical setting of estimating the inverse elasticity of labor supply (e.g., see Sections I and VI of Goldsmith-Pinkham et al. (2020)), the observations are typically clustered at the location level, such as the US commuting zone level, with a short panel dataset of T time periods. The structural equation of interest can thus be written as

$$\tilde{Y}_{g,t} = \tilde{X}_{g,t}\beta + W_{g,t}^\top \gamma + \tilde{e}_{g,t},$$

where g indexes a location, t a time period, $\tilde{Y}_{g,t}$ is wage growth, $\tilde{X}_{g,t}$ is employment growth, and $W_{g,t}$ is a vector of controls which could include location and time fixed effects. Then, according to our notation, G is equal to the number of locations, and $n_g = T$ for all $g \in [G]$. The shift-share IV is an inner product of the initial industry-location shares and the industry-period growth rates, i.e., $\tilde{z}_{g,t} = \sum_{k=1}^{K} \tilde{s}_{0,g,k} h_{k,t}$, where the employment share of industry k in the location g in a certain initial period corresponds to the base IV $\tilde{s}_{0,g,k}$, the growth rate of industry k in the period t corresponds to the weight $h_{k,t}$, and K is the number of industries.

Additionally, let $\tilde{z}$ be the $n \times d_z$-dimensional matrix formed by $\tilde{z}_{i,g}$, and denote $z = M_W \tilde{z}$. In many empirical applications, the dimension d_z is just one (e.g., the shift-share IV), but our setup also allows for $d_z > 1$, while maintaining the requirement that d_z is fixed with respect to the sample size n.

The null and alternative hypotheses studied are $\mathcal{H}_0 : \beta = \beta_0$ against $\mathcal{H}_1 : \beta \neq \beta_0$. We focus on the model with a scalar endogenous variable for two reasons. First, in many empirical applications of IV regressions, there is only one endogenous variable.[5] Second, if we assume that at least the low-dimensional IVs provide strong identification, our results can be extended to testing of scalar restrictions with multiple endogenous variables by applying standard subvector inference methods, without appealing to a projection-based weak-identification-robust inference approach.[6] In addition, our method could potentially be extended to the case in which the identification strength provided by many IVs is mixed for the endogenous variables, by following the approach of Chao et al. (2012). We leave these extensions for future research.

3.2 Test Statistics

Under our setting, it is possible to conduct inference directly based on the low-dimensional IVs. Specifically, given a $d_z \times d_z$ positive definite weighting matrix $\hat{A}_n$, the generalized

[5]For example, 101 out of 230 specifications in Andrews, Stock, and Sun (2019)'s sample and 1,087 out of 1,359 in Young (2022)'s sample feature one endogenous regressor and one IV. Similarly, Lee, McCrary, Moreira, and Porter (2022) find that 61 out of 123 IV papers published in *AER* between 2013 and 2019 use a single IV. Our setting further allows for the overidentified case with one endogenous regressor and multiple IVs. In general, empirical researchers can generate many IVs by using polynomials or interactions based on their low-dimensional base IVs and control variables, in the same spirit as Angrist and Krueger (1991). Then, it is possible to achieve efficiency improvement using our combination procedure.

[6]For weak-identification-robust subvector inference, in general, one may use a projection approach (Dufour and Taamouti, 2005) after implementing inference on the whole vector of endogenous variables. However, the projection approach typically leads to conservative inference. Alternative subvector inference methods for IV regressions (e.g., see Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), Andrews (2017), Guggenberger, Kleibergen, and Mavroeidis (2019, 2021), and Wang and Doko Tchatoka (2018)) provide a power improvement over under a fixed number of instruments (some of these methods further require conditional homoskedasticity). However, whether they can be applied to the setting of many weak instruments is unclear.

method of moments (GMM) estimator can be written as

$$\hat{\beta}_1 = (X^\top z\hat{A}_n z^\top X)^{-1}(X^\top z\hat{A}_n z^\top Y). \tag{3.4}$$

It is also possible to construct test statistics using the K-dimensional many IVs. Denote $P = Z(Z^\top Z)^{-1}Z^\top$ as the projection matrix of Z. Then, the leave-one-cluster-out jackknife IV estimator of β is denoted as $\hat{\beta}_2$ and defined as

$$\begin{aligned}\hat{\beta}_2 &= \left(\sum_{g,h\in[G]^2, g\neq h} X_{[g]}^\top P_{[g,h]} X_{[h]}\right)^{-1}\left(\sum_{g,h\in[G]^2, g\neq h} X_{[g]}^\top P_{[g,h]} Y_{[h]}\right)\\ &= \left(X^\top (P-\bar{P})X\right)^{-1}\left(X^\top (P-\bar{P})Y\right),\end{aligned} \tag{3.5}$$

where $\bar{P}$ is the block diagonal matrix corresponding to P such that the g-th block on its diagonal is $P_{[g,g]}$. Note that under independent data, $\hat{\beta}_2$ reduces to the JIVE estimator in Angrist, Imbens, and Krueger (1999), Chao et al. (2012), and Mikusheva and Sun (2022).

Given $\hat{\beta}_1$ and $\hat{\beta}_2$, we define the estimator $\hat{\beta}$ as

$$\hat{\beta} = \frac{\ddot{\Phi}_2^{1/2}}{\dot{\Phi}_1^{1/2} + \ddot{\Phi}_2^{1/2}} \times \hat{\beta}_1 + \frac{\dot{\Phi}_1^{1/2}}{\dot{\Phi}_1^{1/2} + \ddot{\Phi}_2^{1/2}} \times \hat{\beta}_2, \tag{3.6}$$

where $\dot{\Phi}_1$ and $\ddot{\Phi}_2$ are the variance estimators for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, to be defined later in Section 4.1. We show that $\hat{\beta}$ is consistent whenever either the low-dimensional IV estimator $\hat{\beta}_1$ or the many-IV estimator $\hat{\beta}_2$ is consistent (i.e., either the low-dimensional IVs or the many IVs provide strong identification for β). Because researchers do not need to know which of the two estimators is consistent when constructing $\hat{\beta}$, the estimator is doubly robust.

We then use the doubly robust estimator $\hat{\beta}$ to re-estimate the variances associated with $\hat{\beta}_1$ and $X^\top(P-\bar{P})e$, denoted by $\widehat{\Phi}_1$ and $\hat{\Sigma}$, respectively, and defined in Section 4.1. These variance estimates are used to construct the Wald statistic (with low-dimensional IVs) and the leave-one-cluster-out jackknife LM statistic (with many IVs):

$$T(\beta_0) = \frac{(X^\top z\hat{A}_n z^\top X)^{-1} X^\top z\hat{A}_n z^\top e(\beta_0)}{\sqrt{\widehat{\Phi}_1}} = \frac{\hat{\beta}_1 - \beta_0}{\sqrt{\widehat{\Phi}_1}}, \tag{3.7}$$

$$LM(\beta_0) = \frac{X^\top (P - \bar{P}) e(\beta_0)}{\sqrt{\hat{\Sigma}}}, \tag{3.8}$$

where $e(\beta_0) = Y - X\beta_0$.

Lastly, as pointed out by Hausman et al. (2012), Lim et al. (2024a), and Mikusheva and Sun (2024), it is possible to use the jackknife AR statistic to further improve the efficiency of the jackknife LM statistic. In the current setting with clustered data, we define the leave-one-cluster-out jackknife AR statistic as

$$AR = \frac{\hat{e}^\top (P - \bar{P}) \hat{e}}{\sqrt{\hat{\Upsilon}}}, \tag{3.9}$$

where $\hat{\Upsilon}$ is a consistent variance estimator for the numerator defined later, and $\hat{e} = Y - X\hat{\beta}$.

We use the consistent estimator $\hat{\beta}$ to construct our tests for two reasons. First, using the null value β_0 to form the AR statistic and then combine it with the LM statistic can produce a non-monotonic power curve (i.e., a power ditch against certain alternatives; see Andrews (2016) and Lim et al. (2024a) for related discussions). Constructing the AR statistic with $\hat{\beta}$ avoids this issue. Consequently, the AR statistic in (3.9) does not depend on β_0 but serves as a normalized estimator of zero, used solely to improve the efficiency of our procedure. Second, the combination test relies on correlations between $T(\beta_0)$ and $LM(\beta_0)$ and between $LM(\beta_0)$ and AR, whose consistent estimation requires a consistent β estimator. This consistency is crucial to ensure proper size control, especially when the many-IV specification is weakly identified.

In the next section, we illustrate how to optimally combine the three test statistics $T(\beta_0)$, $LM(\beta_0)$, and AR. The corresponding variance estimators $(\dot{\Phi}_1, \ddot{\Phi}_2, \widehat{\Phi}_1, \hat{\Sigma}, \hat{\Upsilon})$ are introduced later in Section 4.1.

3.3 Combination Test

Given the three test statistics $(T(\beta_0), LM(\beta_0), AR)$, we seek to combine them in a theoretically justified way that can improve on the Wald test based only on the low-dimensional IVs. The key insight of our paper is that under certain local alternative $\beta - \beta_0 = \delta d_n$ with some deterministic sequence $d_n \downarrow 0$, we have the following joint limiting distribution:

$$\begin{pmatrix} T(\beta_0) \\ LM(\beta_0) \\ AR \end{pmatrix} \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} a_1\delta \\ a_2\delta \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_2 \\ 0 & \rho_2 & 1 \end{pmatrix}\right) \tag{3.10}$$

for some a_1, a_2, ρ_1 and ρ_2 to be defined later. In this limiting problem, the UMPU level-α test for the default null hypothesis $\delta = 0$ against two-sided alternatives, which are solely based on the limiting three-dimensional normal random vector, can be obtained by invoking standard hypothesis-testing results (see, for example, Section 4.2 of Lehmann and Romano (2006)), and is stated in Proposition 3.1 below.

Proposition 3.1. *Suppose that one observes $(\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3)$, which follows the limiting distribution in (3.10) with $\rho_1^2 + \rho_2^2 < 1$ and wants to test $\mathcal{H}_0 : \delta = 0$ against $\mathcal{H}_1 : \delta \neq 0$ for known values of $(a_1, a_2, \rho_1, \rho_2)$, then the UMPU level-$\alpha$ test rejects if*

$$\left(\frac{b_1\tilde{\mathcal{N}}_1 + b_2\tilde{\mathcal{N}}_2 + b_3\tilde{\mathcal{N}}_3}{\sqrt{b_1^2 + b_2^2 + b_3^2}}\right)^2 \geqslant \mathbb{C}_\alpha,$$

where $\mathbb{C}_\alpha$ is the $(1-\alpha)$ percentile of a chi-squared random variable with one degree of freedom,

$$\begin{pmatrix} \tilde{\mathcal{N}}_1 \\ \tilde{\mathcal{N}}_2 \\ \tilde{\mathcal{N}}_3 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_2 \\ 0 & \rho_2 & 1 \end{pmatrix}^{-1/2} \begin{pmatrix} \mathcal{N}_1 \\ \mathcal{N}_2 \\ \mathcal{N}_3 \end{pmatrix}, \textit{ and } \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_2 \\ 0 & \rho_2 & 1 \end{pmatrix}^{-1/2} \begin{pmatrix} a_1 \\ a_2 \\ 0 \end{pmatrix}.$$

The corresponding power function for the UMPU test is

$$\mathbb{P}\left(\chi_1^2\left(\delta^2\frac{(1-\rho_2^2)a_1^2-2\rho_1a_1a_2+a_2^2}{1-\rho_1^2-\rho_2^2}\right)\geqslant\mathbb{C}_\alpha\right),$$

where $\chi_1^2(\lambda)$ is a noncentral chi-squared with noncentrality λ and one degree of freedom.

In light of this optimal testing result in the limiting problem, one may wish to propose implementing the following test:

$$\phi_n^o=\mathbf{1}\left\{(\omega_1T(\beta_0)+\omega_2LM(\beta_0)+\omega_3AR)^2\geqslant\mathbb{C}_\alpha\right\}\equiv\phi^*\left(T(\beta_0),LM(\beta_0),AR\right),\tag{3.11}$$

$$\text{where}\quad\begin{pmatrix}\omega_1\\ \omega_2\\ \omega_3\end{pmatrix}=\frac{1}{\sqrt{b_1^2+b_2^2+b_3^2}}\begin{pmatrix}1&\rho_1&0\\ \rho_1&1&\rho_2\\ 0&\rho_2&1\end{pmatrix}^{-1/2}\begin{pmatrix}b_1\\ b_2\\ b_3\end{pmatrix},\tag{3.12}$$

and then investigate its asymptotic justification.

However, the parameters a_1, a_2, ρ_1, ρ_2 are usually unknown and need to be estimated. In addition, it turns out that the weights $(\omega_1, \omega_2, \omega_3)$ are invariant to the scale normalization of (b_1, b_2, b_3), and thus, (a_1, a_2). Therefore, to construct the UMPU test, it suffices to consistently estimate $\alpha_1 = a_1/\sqrt{a_1^2 + a_2^2}$ and $\alpha_2 = a_2/\sqrt{a_1^2 + a_2^2}$ along with ρ_1 and ρ_2.

Given the consistent estimators $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\rho}_1, \hat{\rho}_2)$ for $(\alpha_1, \alpha_2, \rho_1, \rho_2)$ specified in Section 4.1, we then implement the feasible version of the combination test:

$$\phi_n^*=\mathbf{1}\left\{(\hat{\omega}_1T(\beta_0)+\hat{\omega}_2LM(\beta_0)+\hat{\omega}_3AR)^2\geqslant\mathbb{C}_\alpha\right\},\tag{3.13}$$

$$\text{where}\begin{pmatrix}\hat{\omega}_1\\ \hat{\omega}_2\\ \hat{\omega}_3\end{pmatrix}=\frac{1}{\sqrt{\hat{b}_1^2+\hat{b}_2^2+\hat{b}_3^2}}\times\begin{pmatrix}1&\hat{\rho}_1&0\\ \hat{\rho}_1&1&\hat{\rho}_2\\ 0&\hat{\rho}_2&1\end{pmatrix}^{-1/2}\begin{pmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{b}_3\end{pmatrix},\begin{pmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{b}_3\end{pmatrix}=\begin{pmatrix}1&\hat{\rho}_1&0\\ \hat{\rho}_1&1&\hat{\rho}_2\\ 0&\hat{\rho}_2&1\end{pmatrix}^{-1/2}\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\\ 0\end{pmatrix}.\tag{3.14}$$

4 Large-Sample Theory

In this section, we investigate the asymptotic behavior of our combination test. We begin by stating and discussing general assumptions about the data-generating process and the identification strength of both the low-dimensional and many IVs. We then establish the asymptotic efficiency properties of the combination test and, finally, compare its efficiency to that of the conventional Wald test based solely on low-dimensional IVs via the limiting length ratio of their confidence intervals.

4.1 General Assumptions

As in Chao et al. (2012), we treat $\tilde{Z}$ and W as fixed. This is equivalent to treating them as random and repeating all the analyses in the paper by conditioning on them. For the data-generating process, we impose the following assumptions.

Assumption 1. *The following conditions hold when n is sufficiently large:*

1. $\max_{i \in I_g, g \in [G]} \mathbb{E}(\tilde{e}_{i,g}^4 + \tilde{V}_{i,g}^4) \leqslant C < \infty$;

2. $\max_{1 \leqslant g \leqslant G} n_g \leqslant C < \infty$;

3. *Let*

$$\Omega_g = \mathbb{E}\left[\begin{pmatrix} \tilde{e}_{[g]}\tilde{e}_{[g]}^\top & \tilde{e}_{[g]}\tilde{V}_{[g]}^\top \\ \tilde{V}_{[g]}\tilde{e}_{[g]}^\top & \tilde{V}_{[g]}\tilde{V}_{[g]}^\top \end{pmatrix}\right], \quad 1 \leqslant g \leqslant G,$$

then

$$0 < \frac{1}{C} \leqslant \min_{1 \leqslant g \leqslant G} \lambda_{\min}\left(\Omega_g\right) \leqslant \max_{1 \leqslant g \leqslant G} \lambda_{\max}\left(\Omega_g\right) \leqslant C < \infty;$$

4.

$$0 < \frac{1}{C} \leqslant \lambda_{\min}\left(\frac{1}{n} \sum_{i \in I_g, g \in [G]} z_{i,g} z_{i,g}^\top\right) \leqslant \lambda_{\max}\left(\frac{1}{n} \sum_{i \in I_g, g \in [G]} z_{i,g} z_{i,g}^\top\right) \leqslant C < \infty,$$

$$0 < \frac{1}{C} \leqslant \lambda_{\min}\left(\frac{1}{n}\sum_{i\in I_g, g\in[G]} W_{i,g}W_{i,g}^\top\right) \leqslant \lambda_{\max}\left(\frac{1}{n}\sum_{i\in I_g, g\in[G]} W_{i,g}W_{i,g}^\top\right) \leqslant C < \infty,$$

and

$$\max_{i\in I_g, g\in[G]}\left\{\left|\tilde{\Pi}_{i,g}\right| + |\Pi_{i,g}|\right\} \leqslant C < \infty, \quad \max_{i\in I_g, g\in[G]}\left\{\|z_{i,g}\|_2 + \|W_{i,g}\|_2\right\} = o(\sqrt{n});$$

5. *There exists a sequence of non-random positive definite matrices A_n such that*

$$A_n^{-1/2}\hat{A}_n A_n^{-1/2} \xrightarrow{p} I_{d_z},$$

and $\lambda_{\max}(A_n)/\lambda_{\min}(A_n) \leqslant C < \infty$.

Remark 4.1. Assumption 1.1 is a standard condition on the moments of error terms. Assumption 1.2 restricts the cluster size to be bounded, which incorporates cross-sectional and short-panel data structures. It is also possible to extend our analysis to the case with divergent cluster sizes, especially when the within-cluster dependence is weak. However, when the cluster size is allowed to diverge and there is strong within-cluster dependence, the convergence rates of various IV estimators depend on the identification strength, the within-cluster dependence, the cluster sizes, and the number of clusters in a complicated way. We leave the investigation in this direction for future research. Assumption 1.3 ensures that the error covariance matrix is non-singular for each cluster. Assumption 1.4 is a mild condition for the design matrix. Assumption 1.5 states that the weighing matrix $\hat{A}_n$ converges in probability to some non-random positive definite matrix, which is standard in the GMM setup. We note that for two-stage least squares (TSLS), Assumption 1.5 is actually implied by Assumption 1.4.

For the low-dimensional IVs, we focus on the case of strong identification strength in the main text. We further investigate the case in which the low-dimensional IVs have weak identification strength in Supplementary Appendix A, and show that, under certain conditions, when the many IVs provide strong identification, our optimal combination test still asymptotically controls size.

With strong identification, the asymptotic variance of $\hat{\beta}_1$ is

$$\Phi_1 = (\Pi^\top \acute{\Pi})^{-1}\Psi(\Pi^\top \acute{\Pi})^{-1},$$

where $\Omega = \mathbb{E}\sum_{g\in[G]}\left(z_{[g]}^\top \tilde{e}_{[g]}\right)\left(z_{[g]}^\top \tilde{e}_{[g]}\right)^\top$, $\acute{\Pi} = zA_n z^\top \Pi$ and $\Psi = \Pi^\top z A_n \Omega A_n z^\top \Pi$. The cluster-robust variance estimator $\widehat{\Phi}_1$ for the Wald statistic with low-dimensional IVs is

$$\widehat{\Phi}_1 = (X^\top \acute{X})^{-1}\hat{\Psi}(X^\top \acute{X})^{-1}, \tag{4.1}$$

where $\hat{\Omega} = \sum_{g\in[G]}\left(z_{[g]}^\top \hat{e}_{[g]}\right)\left(z_{[g]}^\top \hat{e}_{[g]}\right)^\top$, $\acute{X} = z\hat{A}_n z^\top X$ and $\hat{\Psi} = X^\top z\hat{A}_n\hat{\Omega}\hat{A}_n z^\top X$. The initial estimator $\dot{\Phi}_1$ for Φ_1 used in the computation of $\hat{\beta}$ in (3.6) is defined in the same way as $\widehat{\Phi}_1$, except that $\hat{e}_{[g]} = Y_{[g]} - X_{[g]}\hat{\beta}$ is replaced by $\dot{e}_{[g]} = Y_{[g]} - X_{[g]}\hat{\beta}_1$.

We make the following assumptions regarding the inference with low-dimensional IVs.

Assumption 2. *Let $r_n = \left\|z^\top \Pi\right\|_2$, then $\sqrt{n}/r_n \to 0$.*

Remark 4.2. Assumption 2 ensures that $z^\top \Pi$, the deterministic component of $z^\top X$, dominates $z^\top V$, its random component. This condition is therefore key for the consistency of $\hat{\beta}_1$ for β, that is, it ensures strong identification by the low-dimensional IVs. Throughout the paper, we focus on this strong identification case for the low-dimensional IVs while allowing the many-IV specification to be either weakly or strongly identified. In Section 4.2, we show that the optimal combination test adapts to the identification strength of the many IVs. In particular, it controls size asymptotically and retains nontrivial power even when $\hat{\beta}_2$ is inconsistent. In Section A of the Appendix, we consider the opposite scenario in which the low-dimensional IV specification is weakly identified while the many-IV specification is strongly identified. We provide mild conditions under which the exact inference procedure proposed in the paper still controls asymptotic size under the null.

If the many-IV-based identification is strong, similar to Chao et al. (2012), we can show that the asymptotic variance of $\hat{\beta}_2$ is

$$\Phi_2 = \left(\Pi^\top (P - \bar{P})\Pi\right)^{-1} \Sigma \left(\Pi^\top (P - \bar{P})\Pi\right)^{-1},$$

where

$$\Sigma = \mathbb{E}\left(\sum_{g,h\in[G]^2, g\neq h} \Pi_{[g]}^\top P_{[g,h]} \left(\sum_{k\in[G]} M_{W,[h,k]}\tilde{e}_{[k]}\right)\right)^2 + \mathbb{E}\left(\sum_{g,h\in[G]^2, g\neq h} \tilde{V}_{[g]}^\top P_{[g,h]}\tilde{e}_{[h]}\right)^2$$

is the asymptotic variance of $X^\top(P - \bar{P})e$. A natural estimator for Φ_2 is thus

$$\widehat{\Phi}_2 = \left(X^\top (P - \bar{P})X\right)^{-1} \hat{\Sigma} \left(X^\top (P - \bar{P})X\right)^{-1}, \tag{4.2}$$

where $\hat{\Sigma}$ is a consistent estimator of Σ. Such an estimator is proposed in Chao et al. (2012) for the case with independent data. Here, in addition to extending to clustered data, we need to account for the fact that W has already been partialled out, whereas in Chao et al. (2012) the coefficients for W are also estimated. Therefore, some adjustments are required, as in Matsushita and Otsu (2024). For that purpose, define $Q = M_W(P - \bar{P})M_W$, and let $\bar{Q}$ be the block diagonal matrix corresponding to Q such that the g-th block on its diagonal is $Q_{[g,g]}$. Our variance estimator is similar to the one in Chao et al. (2012) but with $P - \bar{P}$ replaced by $Q - \bar{Q}$, i.e.,

$$\hat{\Sigma} = \sum_{g\in[G]}\left(\sum_{h\in[G], h\neq g} \tilde{X}_{[h]}^\top Q_{[h,g]}\hat{e}_{[g]}\right)^2 + \sum_{g,h\in[G]^2, g\neq h}\left(\tilde{X}_{[g]}^\top Q_{[g,h]}\hat{e}_{[h]}\right)\left(\tilde{X}_{[h]}^\top Q_{[h,g]}\hat{e}_{[g]}\right). \tag{4.3}$$

The initial estimator $\ddot{\Phi}_2$ for Φ_2 used in the computation of $\hat{\beta}$ is defined in the same way as $\widehat{\Phi}_2$, except that $\hat{e}_{[g]} = Y_{[g]} - X_{[g]}\hat{\beta}$ is replaced by $\ddot{e}_{[g]} = Y_{[g]} - X_{[g]}\hat{\beta}_2$.

Last, for the jackknife AR statistic, the variance estimator is given by

$$\hat{\Upsilon} = 2 \sum_{g,h \in [G]^2, g \neq h} \left(\hat{e}_{[g]}^\top P_{[g,h]} \hat{e}_{[h]} \right)^2, \tag{4.4}$$

which is consistent for the asymptotic variance of $\hat{e}^\top (P - \bar{P}) \hat{e}$, given by

$$\Upsilon = \mathbb{E} \left(\sum_{g,h \in [G]^2, g \neq h} \tilde{e}_{[g]}^\top P_{[g,h]} \tilde{e}_{[h]} \right)^2.$$

We make the following assumptions regarding the inference with many IVs.

Assumption 3. *1. $K \to \infty$ as $n \to \infty$ such that $\limsup_{n \to \infty} K/n \leqslant C < 1$;*

2. $rank(P) = K$ and $\max_{1 \leqslant g \leqslant G} \lambda_{\max} \left(P_{[g,g]} \right) \leqslant C < 1$;

3. Let $\hat{\Pi} = M_W (P - \bar{P}) \Pi = Q \tilde{\Pi}$ and $\bar{\Pi} = (Q - \bar{Q}) \tilde{\Pi}$, then

$$\max_{i \in I_g, g \in [G]} \left\{ \left| \hat{\Pi}_{i,g} \right| + \left| \bar{\Pi}_{i,g} \right| \right\} \leqslant C < \infty,$$

and

$$\tilde{\Pi}^\top \tilde{\Pi} \leqslant C \Pi^\top \Pi, \;\; \hat{\Pi}^\top \hat{\Pi} \geqslant \Pi^\top \Pi / C, \;\; \Pi^\top (P - \bar{P}) \Pi \geqslant \Pi^\top \Pi / C,$$

when n is large enough;

4. For all sufficiently large n,

$$\left| corr \left(\sum_{g,h \in [G]^2, g \neq h} \tilde{V}_{[g]}^\top P_{[g,h]} \tilde{V}_{[h]}, \sum_{g,h \in [G]^2, g \neq h} \tilde{V}_{[g]}^\top P_{[g,h]} \tilde{e}_{[h]} \right) \right| \leqslant C < 1.$$

Remark 4.3. Assumption 3.1 allows the dimension of many IVs K to be proportional to the sample size n. Assumption 3.2 is similar to the standard condition that $\max_{1 \leqslant i \leqslant n} P_{ii} \leqslant C < 1$

in the literature on many instruments, and the restriction that $\text{rank}(P) = K$ will exclude redundant columns from Z. Assumption 3.3 holds in general if $\tilde{\Pi}^\top\tilde{\Pi}$, $\Pi^\top\Pi$ and $\Pi^\top P\Pi$ are of the same order. Assumption 3.4 excludes perfect correlations between the two quadratic forms. Note that we do not impose any restriction on the identification strength of many IVs, i.e., we allow $\Pi^\top\Pi/\sqrt{K}$ to be bounded.

4.2 Asymptotic Efficiency Properties of Combination Test

We now investigate the asymptotical properties of ϕ_n^* when the low-dimensional IVs are strong, in the sense that Assumption 2 holds. This allows us to define the local alternative according to the asymptotic variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the limiting covariance structure of the component statistics, from which the joint limiting distribution of $(T(\beta_0), LM(\beta_0), AR)$ can be derived. The results with weak low-dimensional IVs are given in Supplementary Appendix A. The formal regularity condition is stated as follows.

Assumption 4. *The following limits exist:*

$$\rho_1 = \lim_{n\to\infty} \frac{1}{\sqrt{\Psi\Sigma}} \sum_{g\in[G]} \mathbb{E}\left[\left(\acute{\Pi}_{[g]}^\top \tilde{e}_{[g]}\right)\left(\hat{\Pi}_{[g]}^\top \tilde{e}_{[g]}\right)\right],$$

$$\rho_2 = \lim_{n\to\infty} \frac{2}{\sqrt{\Sigma\Upsilon}} \sum_{g,h\in[G]^2, g\neq h} \mathbb{E}\left[\left(\tilde{V}_{[g]}^\top P_{[g,h]}\tilde{e}_{[h]}\right)\left(\tilde{e}_{[g]}^\top P_{[g,h]}\tilde{e}_{[h]}\right)\right],$$

with $\rho_1^2 + \rho_2^2 < 1$.

Remark 4.4. Under many instruments, the asymptotic expansion of $LM(\beta_0)$ includes both linear and quadratic functions of the errors $(\tilde{e}, \tilde{V})$, whereas AR depends only on a quadratic function of $\tilde{e}$. The linear and quadratic components are asymptotically normal and uncorrelated, and therefore asymptotically independent. Since the Wald statistic $T(\beta_0)$ involves only linear functions of the errors, it is asymptotically uncorrelated with AR. Finally, ρ_1 and ρ_2 denote, respectively, the correlation between the linear components of $T(\beta_0)$ and $LM(\beta_0)$, and the correlation between the quadratic components of $LM(\beta_0)$ and AR.

The following theorem establishes the joint distribution of the three test statistics above under the local alternative.

Theorem 4.1. *Under Assumptions 1–4, suppose that there exists a deterministic sequence $d_n \downarrow 0$ such that $d_n \Phi_1^{-1/2} \to a_1$, $d_n \Phi_2^{-1/2} \to a_2$, and $\beta - \beta_0 = \delta d_n$ for some fixed δ, where $a_1 \geqslant 0$, $a_2 \geqslant 0$ and $a_1^2 + a_2^2 > 0$. Then we have the joint limiting distribution (3.10) for $(T(\beta_0), LM(\beta_0), AR)^\top$.*

Remark 4.5. The existence of the sequence d_n is ensured by Assumption 2. In particular, we may define $d_n = \min\left(\Phi_1^{1/2},\, \Phi_2^{1/2}\right)$. Under the strong identification of the low-dimensional IVs in Assumption 2, we have $\Phi_1^{1/2} = O\left(\frac{\sqrt{n}}{r_n}\right) = o(1)$, which immediately implies that $d_n = \min\left(\Phi_1^{1/2},\, \Phi_2^{1/2}\right) = o(1)$, regardless of the order of Φ_2.

Remark 4.6. The joint normality established in Theorem 4.1 holds even when the many-IV specification is weakly identified in the sense of Mikusheva and Sun (2022), that is, when $\Pi^\top \Pi / \sqrt{K}$ is bounded. This result follows from two observations. First, the estimator $\hat{\beta}$ used to construct the variance estimators $\hat{\Phi}_2$ and $\hat{\Upsilon}$ for the LM and AR statistics remains consistent due to the strong identification of the low-dimensional IVs and the double robustness of $\hat{\beta}$. Second, under weak many-IV identification, the quadratic components of the LM and AR statistics dominate their asymptotic behavior and yield asymptotic normality as long as $K \to \infty$. In this regime, we have $\rho_1 = 0$, $d_n = \sqrt{n}/r_n$, and $\Phi_2^{-1} = O(1)$, which further imply that $a_2 = 0$ under Assumption 2.

To implement the optimal test ϕ_n^* defined in (3.13), we still need to estimate α_1, α_2, ρ_1, and ρ_2. For α_1 and α_2, we propose the following estimators:

$$\hat{\alpha}_1 = \frac{\sqrt{\widehat{\Phi}_2}}{\sqrt{\widehat{\Phi}_1 + \widehat{\Phi}_2}}, \quad \text{and} \quad \hat{\alpha}_2 = \frac{\sqrt{\widehat{\Phi}_1}}{\sqrt{\widehat{\Phi}_1 + \widehat{\Phi}_2}}. \tag{4.5}$$

In addition, let $\hat{X} = M_W(P - \bar{P})X$. For ρ_1 and ρ_2, we propose the following estimators:

$$\hat{\rho}_1 = \frac{1}{\sqrt{\hat{\Psi}\hat{\Sigma}}} \sum_{g \in [G]} \left[\left(\acute{X}_{[g]}^\top \hat{e}_{[g]} \right) \left(\hat{X}_{[g]}^\top \hat{e}_{[g]} \right) \right], \text{ and} \tag{4.6}$$

$$\hat{\rho}_2 = \frac{2}{\sqrt{\hat{\Sigma}\hat{\Upsilon}}} \sum_{g,h \in [G]^2, g \neq h} \left[\left(X_{[g]}^\top P_{[g,h]} \hat{e}_{[h]} \right) \left(\hat{e}_{[g]}^\top P_{[g,h]} \hat{e}_{[h]} \right) \right]. \tag{4.7}$$

By combining Proposition 3.1 and Theorem 4.1, and invoking the approach developed by Müller (2011), we obtain a precise sense of asymptotic optimality for our proposed test ϕ_n^*, which is formalized in the following Theorem 4.2.

Theorem 4.2. *Let $\mathcal{M}$ denote the set of data generating processes m that satisfy the conditions of Theorem 4.1 pointwise for all $\delta \in \Re$. Suppose that one wants to test $\mathcal{H}_0 : \delta = 0$ against $\mathcal{H}_1 : \delta \neq 0$. Then, for the class $\mathfrak{C}$ of tests ϕ_n satisfying that*

$$\lim_{n \to \infty} \mathbb{E}\left[\phi_n\right] \leqslant \alpha \quad \textit{for all } m \in \mathcal{M}, \delta = 0, \tag{4.8}$$

$$\liminf_{n \to \infty} \mathbb{E}\left[\phi_n\right] \geqslant \alpha \quad \textit{for all } m \in \mathcal{M}, \delta \neq 0, \tag{4.9}$$

we have $\phi_n^* \in \mathfrak{C}$, and, for any $\delta_1 \neq 0$ and any $\phi_n \in \mathfrak{C}$,

$$\lim_{n \to \infty} \mathbb{E}\left[\phi_n\right] \leqslant \lim_{n \to \infty} \mathbb{E}\left[\phi_n^*\right] \quad \textit{for all } m \in \mathcal{M}, \delta = \delta_1. \tag{4.10}$$

Moreover, for the test $\tilde{\phi}_n = \mathbf{1}\{T^2(\beta_0) \geqslant \mathbb{C}_\alpha\}$, we have $\tilde{\phi}_n \in \mathfrak{C}$, and for any δ and all $m \in \mathcal{M}$,

$$\lim_{n \to \infty} \mathbb{E}\left[\tilde{\phi}_n\right] = \lim_{n \to \infty} \mathbb{E}\left[\phi_n^*\right] \quad \textit{if and only if} \quad a_2 = \rho_1 a_1.$$

Remark 4.7. Theorem 4.2 shows that, under local alternatives, ϕ_n^* attains the asymptotic efficiency bound within the class of tests that remain asymptotically unbiased and valid for all data generating processes inducing the same weak limit for $(T(\beta_0), LM(\beta_0), AR)^\top$ as in

Theorem 4.1. This class of tests includes, in particular, the Wald and jackknife LM tests based solely on the low-dimensional and many IVs, respectively. It also includes the HLIM and HFUL-based tests of Hausman et al. (2012), both of which are asymptotically equivalent to linear combinations of the LM and AR tests under many IVs.[7] Furthermore, as we will discuss more in detail in Section 4.3, the weights in (3.12) also yield an optimally efficient combined estimator of $(\hat{\beta}_1, \hat{\beta}_2, AR)$ that achieves the minimal asymptotic variance.

However, our optimal test ϕ_n^* does not dominate tests that cannot be expressed directly as functions of $(T(\beta_0), LM(\beta_0), AR)$, such as the sup-score test of Belloni et al. (2012) and the ridge-regularized AR test of Dovì, Kock, and Mavroeidis (2024). Indeed, it is possible to construct data generating processes under which either the optimal combination test, the sup-score test, or the ridge-regularized AR test achieves the highest power. We establish the notion of optimality in Theorem 4.2 primarily to provide guidance for constructing tests that improve upon the conventional Wald test, rather than to identify a globally optimal procedure. That said, one could potentially combine our ϕ_n^* test with alternative tests such as the sup-score test, in the spirit of Navjeevan (2024), to obtain more powerful inference.

Remark 4.8. Theorem 4.2 also clarifies the necessary and sufficient condition under which the combination test does not deliver a strict power gain over the Wald test $T(\beta_0)$, namely $a_2 = \rho_1 a_1$. Recall that a_1 and a_2 represent the orders of the concentration parameters for the low-dimensional and many IVs, respectively. As shown below, even when we allow either of them to be zero—thereby covering situations where one IV estimator dominates the other in terms of convergence rate—the condition ($a_2 = \rho_1 a_1$) is still seldom met. Put differently, one should generally anticipate strictly more powerful inference when using our test.

In particular, when $a_1 = 0$ and $a_2 > 0$, corresponding to the case when the identification strength of many IVs is larger than that of the low-dimensional IVs (i.e., the convergence rate of $\hat{\beta}_2$ is faster than that of $\hat{\beta}_1$), we obtain a strict power improvement for all values of ρ_1 and ρ_2 satisfying $\rho_1^2 + \rho_2^2 < 1$. Conversely, when $a_1 > 0$ and $a_2 = 0$, meaning that low-dimensional

[7] E.g., see the discussions below Theorem 4.2 in Lim et al. (2024a).

IVs provide stronger identification than many IVs, we still achieve a strict power gain as long as $\rho_1 \neq 0$. When $a_1 > 0$ and $a_2 > 0$, that is, when the two sets of IVs have identification strengths of the same order, strict power improvement is ensured, provided that $\rho_1 \neq a_2/a_1$. Indeed, at $\rho_1 = a_2/a_1$, the sufficient statistic for δ derived from the joint limiting distribution of the three component statistics (cf. (3.10)) becomes independent of the limiting Gaussian observations associated with the LM and AR statistics, so it is not surprising that combining the Wald statistic with them does not yield a more powerful inference.

Remark 4.9. As long as the low-dimensional IVs provide strong identification, the optimal combination test ϕ_n^* does not lose asymptotic power for any degree of identification strength of the many IVs. In this sense, the efficiency gains delivered by the combination test are essentially a "free lunch". Under local alternatives, the weak convergence result in Theorem 4.1 holds uniformly, regardless of whether the many IVs are strong or weak. In particular, when the many IVs are weak so that $a_2 = \rho_1 = 0$ (as the quadratic term in $LM(\beta_0)$ dominates the linear term), the second part of Theorem 4.2 shows that the combination test asymptotically reduces to the Wald test, implying no efficiency loss from combining. Moreover, as shown below, the combination test remains consistent against fixed alternatives irrespective of the identification strength of the many IVs.

Finally, for any fixed alternative, both $T(\beta_0)$ and $LM(\beta_0)$ are consistent and, by construction, avoid the issue of non-monotonic power (noted at the end of Section 3.2) when their corresponding set of IVs is strong. Hence, it is reasonable to anticipate that our combined test will retain these desirable properties, a result that we formalize in the theorem below. However, we emphasize once more that these results remain valid regardless of the strength of the many IVs.

Theorem 4.3. *Suppose that Assumptions 1-4 hold. Then, under $\beta - \beta_0 = \delta$ for some fixed $\delta \neq 0$, we have $\lim_{n\to\infty} \mathbb{E}\left[\phi_n^*\right] = 1$.*

4.3 Quantifying Efficiency Improvement in Large Samples

We measure the efficiency improvement of the combination test over the conventional Wald test that uses only low-dimensional IVs by the percentage reduction in the asymptotic length of the resulting confidence interval. Recall from Remark 4.9 that, under weak many instruments, the combination test is asymptotically equivalent to the Wald test. Consequently, the associated confidence interval takes the usual "estimator plus and minus a standard error times a critical value" form: $\big[\hat{\beta}_1 - \sqrt{\widehat{\Phi}_1} \times \sqrt{\mathbb{C}_\alpha},\ \hat{\beta}_1 + \sqrt{\widehat{\Phi}_1} \times \sqrt{\mathbb{C}_\alpha}\big]$, where $\hat{\beta}_1$ and $\hat{\Phi}_1$ are as defined above, and $\sqrt{\mathbb{C}_\alpha}$ is the standard normal critical value. In this case, the asymptotic efficiency gain over the conventional Wald test is zero.

When the many IVs provide strong identification, the confidence interval associated with our combination test can also be expressed in the familiar "estimator plus and minus a standard error times a critical value" form. To see this, observe that under strong identification by many IVs, the component LM statistic can be represented as follows:

$$LM(\beta_0) = \frac{X^\top (P - \bar{P}) e(\beta_0)}{\sqrt{\hat{\Sigma}}} = \frac{\hat{\beta}_2 - \beta_0}{\sqrt{\widehat{\Phi}_2}} + o_P(1),$$

where the $o_P(1)$ term comes from the fact that under strong identification,

$$\operatorname{sign}(X^\top (P - \bar{P}) X) \xrightarrow{p} 1.$$

Inserting this into (3.13) gives the following form of our combination test:

$$\phi_n^* = \mathbf{1}\left\{ \left(\hat{\omega}_1 \frac{\hat{\beta}_1 - \beta_0}{\sqrt{\widehat{\Phi}_1}} + \hat{\omega}_2 \frac{\hat{\beta}_2 - \beta_0}{\sqrt{\widehat{\Phi}_2}} + \hat{\omega}_2 o_P(1) + \hat{\omega}_3 AR \right)^2 \geqslant \mathbb{C}_\alpha \right\}.$$

The resulting confidence interval is asymptotically equivalent to

$$CI^* = \left[\hat{\beta}^* - \frac{1}{\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1} + \hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}\sqrt{\mathbb{C}_\alpha},\ \hat{\beta}^* + \frac{1}{\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1} + \hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}\sqrt{\mathbb{C}_\alpha} \right], \quad (4.11)$$

where $\hat{\beta}^*$ is a combined estimator of β,

$$\hat{\beta}^* = \frac{\hat{\omega}_1/\sqrt{\widehat{\Phi}_1}}{\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1} + \hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}\hat{\beta}_1 + \frac{\hat{\omega}_2/\sqrt{\widehat{\Phi}_2}}{\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1} + \hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}\hat{\beta}_2 + \frac{\hat{\omega}_3 AR}{\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1} + \hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}.$$

The intuition behind the combined estimator is fundamentally efficiency-driven. First, $\widehat{\Phi}_1$ and $\widehat{\Phi}_2$ estimate the asymptotic variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, so for given weights $(\hat{\omega}_1, \hat{\omega}_2)$ the combined estimator $\hat{\beta}^*$ assigns greater weight to the estimator with the smaller asymptotic variance. Second, although the AR statistic is asymptotically centered at zero and therefore does not affect the location of the combined estimator, its correlation with the many-IV-based estimator $\hat{\beta}_2$ allows it to reduce the variance of the combined estimator. This mechanism parallels the construction of the HLIML and HFUL estimators in Hausman et al. (2012), which can be more efficient than the JIVE estimator. Third, the estimated weights $(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3)$ are consistent for the population weights $(\omega_1, \omega_2, \omega_3)$ in (3.12) associated with the UMPU test. Finally, the optimal weights defined in (3.12) solve the following problem:

$$\min_{\omega_1,\omega_2,\omega_3} \frac{1}{(a_1\omega_1 + a_2\omega_2)^2} \quad \text{s.t.} \quad (\omega_1, \omega_2, \omega_3)\begin{pmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_2 \\ 0 & \rho_2 & 1 \end{pmatrix}\begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = 1.$$

The quadratic constraint ensures that CI^* attains the correct asymptotic coverage, while the objective corresponds to the asymptotic variance of the combined estimator $\hat{\beta}^*$, since

$$\frac{1}{d_n^2\left(\omega_1/\sqrt{\Phi_1}+\omega_2/\sqrt{\Phi_2}\right)^2} \to \frac{1}{(a_1\omega_1+a_2\omega_2)^2}.$$

Thus, the optimal weights in (3.12), originally motivated by the UMPU testing problem, also yield an optimally efficient combined estimator of $(\hat{\beta}_1, \hat{\beta}_2, AR)$ that achieves the minimal asymptotic variance.

A direct implication of the confidence interval in (4.11) is that, asymptotically, the percentage reduction in its length relative to the confidence interval based on the conventional Wald test can be derived analytically as follows,

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty} 1-\frac{1/\left(\hat{\omega}_1/\sqrt{\widehat{\Phi}_1}+\hat{\omega}_2/\sqrt{\widehat{\Phi}_2}\right)}{\sqrt{\widehat{\Phi}_1}} \\
=&1-\sqrt{\frac{(1-\rho_1^2-\rho_2^2)}{(1-\rho_2^2)-2\rho_1 a_2/a_1+(a_2/a_1)^2}} &(4.12)\\
\geqslant&1-\sqrt{\frac{(1-\rho_1^2)a_1^2}{a_1^2-2\rho_1 a_1 a_2+a_2^2}}=1-\sqrt{\frac{(1-\rho_1^2)}{(1-\rho_1^2)+(\rho_1-a_2/a_1)^2}}. &(4.13)
\end{aligned}$$

As equation (4.12) shows, the efficiency gain is primarily driven by the relative identification strength of the low-dimensional and many IVs, summarized by

$$\frac{a_2}{a_1}=\lim_{n\to\infty}\sqrt{\frac{\Phi_1}{\Phi_2}},$$

as well as by the limiting correlations between the Wald and LM statistics, denoted by ρ_1, and between the LM and AR statistics, denoted by ρ_2. In particular, consistent with Theorem 4.2, when $a_2/a_1=\rho_1$, the combination test ϕ_n^* yields no efficiency improvement, and its confidence interval is asymptotically of the same length as that based on $\tilde{\phi}_n$. This case includes the weak many-IV scenario, where $\rho_1=a_2=0$.

By contrast, whenever $a_2/a_1\neq\rho_1$, the resulting confidence interval is strictly shorter, implying improved efficiency. Moreover, the efficiency gain in (4.12) is monotonically increasing in $|\rho_2|$, which leads to the lower bound reported in (4.13) by setting $\rho_2 = 0$. This lower bound can also be interpreted as the CI length reduction achieved by optimally combine $\hat{\beta}_1$ and $\hat{\beta}_2$ only.[8]

Figure 1 plots this bound as a function of a_2/a_1, the ratio of the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$, for various values of ρ_1.

In practice, a_2/a_1 can be estimated by the ratio of any consistent estimator of Φ_1 and Φ_2 under strong identification.[9] This yields a simple rule of thumb, discussed in Section 2: for empirically plausible values of ρ_1 between -0.7 and 0.7, if the ratio of the reported standard errors $\sqrt{\dot{\Phi}_1}$ and $\sqrt{\ddot{\Phi}_2}$ exceeds 1.05, then the associated confidence interval shortens by at least 10%.

In some applications (e.g., the replication of Card (2009) by Goldsmith-Pinkham et al. (2020)), researchers report alternative many-IV estimators (denoted $\tilde{\beta}_2$ with variance $\tilde{\Phi}_2$), such as HFUL, instead of the JIVE estimator ($\hat{\beta}_2$) considered above. Let $\tilde{\rho}_1$ denote the correlation between the GMM estimator $\hat{\beta}_1$ and $\tilde{\beta}_2$. By the same argument, if $\tilde{\rho}_1 \in [-0.7, 0.7]$

[8] The optimal reduction in this case is $\lim_{n\to\infty} 1 - \frac{1/\left(\tilde{\omega}_1/\sqrt{\Phi_1} + \tilde{\omega}_2/\sqrt{\Phi_2}\right)}{\sqrt{\Phi_1}} = 1 - \frac{1}{(\tilde{\omega}_1 + \tilde{\omega}_2 a_2/a_1)}$, where the optimal weights are computed by

$$(\tilde{\omega}_1, \tilde{\omega}_2) = \arg\min_{\omega_1,\omega_2} \frac{1}{(a_1\omega_1 + a_2\omega_2)^2} \quad \text{s.t.} \quad (\omega_1, \omega_2)\begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = 1.$$

By direct calculation, we can show that

$$1 - \frac{1}{(\tilde{\omega}_1 + \tilde{\omega}_2 a_2/a_1)} = 1 - \sqrt{\frac{(1-\rho_1^2)}{(1-\rho_1^2) + (\rho_1 - a_2/a_1)^2}}.$$

[9] If the many-IV specification is weakly identified, then $\Phi_1 = o(1)$ and $1/\Phi_2 = O(1)$, so that $a_2/a_1 = 0$. Suppose we estimate Φ_1 and Φ_2 by $\dot{\Phi}_1$ and $\ddot{\Phi}_2$, respectively. Although $\ddot{\Phi}_2$ is not consistent under weak many IVs, Section C.8 shows that $\ddot{\Phi}_2^{-1/2} = O_P(1)$ and $\dot{\Phi}_1^{1/2} = o_P(1)$, implying $\sqrt{\dot{\Phi}_1/\ddot{\Phi}_2} \xrightarrow{p} 0 = a_2/a_1$. Hence, the plug-in estimator of a_2/a_1 remains consistent even under weak identification of the many-IV specification.

(or $\tilde{\rho}_1 \in [-0.99, 0.99]$), the confidence interval $\widetilde{CI}$ based on the optimal combination of $\hat{\beta}_1$ and $\tilde{\beta}_2$ is at least 10% shorter than the conventional Wald interval whenever the ratio $\sqrt{\Phi_1/\tilde{\Phi}_2}$ exceeds 1.05 (or 1.1). Moreover, if $\tilde{\beta}_2$ is asymptotically equivalent to a combination of the JIVE estimator $\hat{\beta}_2$ and the AR statistic, as is the case for HLIML and HFUL, then our optimal confidence interval (CI^*) is weakly shorter than $\widetilde{CI}$ and therefore at least 10% shorter than the conventional Wald interval. Hence, the rule of thumb is not specific to the JIVE estimator $\hat{\beta}_2$, but applies more broadly to many-IV estimators asymptotically equivalent to a combination of JIVE and the AR statistic.

5 Practical Implementation and Simulation Study

In this section, we first synthesize the preceding discussions to describe how to practically implement our combination test in an empirically relevant model and then apply it in simulations to assess its finite-sample power performance. In particular, we consider the following model with clustered data,

$$\bar{Y}_{i,g} = \bar{X}_{i,g}\beta + \bar{W}_{i,g}^{\top}\gamma + \alpha_g + \bar{e}_{i,g}, \tag{5.1}$$

$$\bar{X}_{i,g} = \bar{Z}_{i,g}^{\top}\pi + \bar{W}_{i,g}^{\top}\tau + \xi_g + \bar{V}_{i,g}, \tag{5.2}$$

where α_g and ξ_g denote cluster-specific fixed effects, and $\bar{X}_{i,g}$, $\bar{W}_{i,g}$, and $\bar{Z}_{i,g}$ represent, respectively, the endogenous regressor, exogenous regressors, and (potentially many) base instruments. By demeaning at the cluster level, we partial out the fixed effects and obtain $\tilde{Y}_{i,g}$, $\tilde{X}_{i,g}$, $W_{i,g}$, $\tilde{Z}_{i,g}$, $\tilde{V}_{i,g}$, and $\tilde{e}_{i,g}$, echoing the notation in our initial model setup in (3.1)-(3.2). The setting we consider in this paper is one in which practitioners are often interested in testing $\mathcal{H}_0 : \beta = \beta_0$ against $\mathcal{H}_1 : \beta \neq \beta_0$ using an alternative set of low-dimensional IVs, $\tilde{z}_{i,g} = f_{i,g}(\tilde{Z}, W)$. These instruments can be formed, for example, by selecting a subset of the original base IVs $\tilde{Z}_{i,g}$, or by taking a weighted or simply an unweighted average of them.

5.1 Practical Implementation

Our implementation procedure starts by partialling out the covariates $\tilde{W}$ from $\tilde{Y}$, $\tilde{X}$, $\tilde{Z}$, and $\tilde{z}$. This produces Y, X, Z, and z, which are consistent with the notation used in Section 3.1. These transformed variables serve as the effective observations in all subsequent estimation and inference procedures. We outline these subsequent steps below in sequential order.

1. *Preliminary estimates of β.*

   (a) Use Y, X, the low-dimensional IVs z, and a given weighting matrix $\hat{A}_n$ to obtain a preliminary GMM estimate $\hat{\beta}_1$ by (3.4);

   (b) Use Y, X, base IVs Z, and the projection matrix of Z to obtain a preliminary leave-one-cluster-out jackknife estimate $\hat{\beta}_2$ by (3.5).

2. *Preliminary variance estimates.* (See Section 4.1 for explicit formulas of the asymptotic variances Φ_1 and Φ_2, and their corresponding estimators.)

   (a) Using the residuals from the structural equation (3.3), computed with $\hat{\beta}_1$, obtain a preliminary estimate $\dot{\Phi}_1$ for the asymptotic variance Φ_1 of $\hat{\beta}_1$.

   (b) Using the residuals from the structural equation (3.3), computed with $\hat{\beta}_2$, obtain a preliminary estimate $\ddot{\Phi}_2$ for the asymptotic variance Φ_2 of $\hat{\beta}_2$. Pay special attention to the adjustments made in (4.3) to account for the fact that W has already been partialled out.

3. *A doubly robust estimate of β and the resulting robust variance estimates.* (See Section 4.1 for explicit formulas of $\hat{\Psi}$, $\hat{\Omega}$, $\widehat{\Phi}_1$, $\hat{\Sigma}$, $\widehat{\Phi}_2$ and $\hat{\Upsilon}$.)

   (a) Use $\hat{\beta}_1$, $\hat{\beta}_2$, $\dot{\Phi}_1$, and $\ddot{\Phi}_2$ to obtain a doubly robust estimate $\hat{\beta}$ of β by (3.6).

   (b) Using the residuals from (3.3), computed with $\hat{\beta}$, obtain the robust variance estimates $\hat{\Psi}$, $\hat{\Omega}$ and thus $\widehat{\Phi}_1$ (cf. (4.1)), $\hat{\Sigma}$ (cf. (4.3)) and thus $\widehat{\Phi}_2$ (cf. (4.2)), and finally $\hat{\Upsilon}$ (cf. (4.4)).

4. *Component statistics and their estimated weights.*

   (a) Using the hypothesized β_0 and the outputs from Step 3 above, compute the three component statistics—$T(\beta_0)$ (cf. (3.7),) $LM(\beta_0)$ (cf. (3.8)) and AR (cf. (3.9)).

   (b) Using the outputs from Step 3 above, compute $\hat{\alpha}_1$ and $\hat{\alpha}_2$ (cf. (4.5)) and $\hat{\rho}_1$ and $\hat{\rho}_2$ (cf. (4.6) and (4.7)), and thus the weights $(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3)$ (cf. (3.14)).

5. Using outputs from Step 4, obtain the feasible combination test ϕ_n^* by (3.13).

5.2 Simulation Study

All of our simulations are based on (5.1) and (5.2). The fixed effects are generated by $\alpha_g = u_{1g} + g/G$ and $\xi_g = u_{2g} + g/G$, $g = 1, \ldots, G$, where u_{1g} and u_{2g} are independent standard normal random variables. The control variables in $\bar{W}_{i,g}$ are generated by the standard normal distribution, and the dimension of $\bar{W}$ is fixed at $d_w = 10$. The instruments in $\bar{Z}_{i,g}$ are normally distributed with mean 0 and cluster-level dependence: within each cluster g, the covariance matrix is given by

$$\Omega_{1g} = \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_1 \\ \theta_1 & 1 & \cdots & \theta_1 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_1 & \theta_1 & \cdots & 1 \end{bmatrix}_{n_g \times n_g}, \quad g = 1, \ldots, G,$$

and between clusters these instruments are independent of each other; we set $\theta_1 = 0.5$ in our simulations. Finally, to obtain an arguably complex error structure, we first generate $\acute{e}_{i,g} = \rho\varepsilon_{i,g} + \sqrt{1-\rho^2}\sigma_{i,g}v_g$ and $\acute{V}_{i,g} = \rho\eta_{i,g} + \sqrt{1-\rho^2}\sigma_{i,g}v_g$, where $\sigma_{i,g} = \sqrt{\left(0.2 + (\bar{W}_{i,g}^\top \tau)^2\right)/2.4}$. Here, $\varepsilon_{i,g}$, $\eta_{i,g}$, and v_g are mutually independent standard normal random variables, ρ governs the degree of endogeneity, τ is specified below, and we fix $\rho = 0.5$ in all simulations. Next, within each cluster, we premultiply the vectors $\acute{e}_{i,g}$ and $\acute{V}_{i,g}$ by

$$\Omega_{2g} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \theta_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta_2^{n_g-1} & \theta_2^{n_g-2} & \cdots & 1 \end{bmatrix}_{n_g \times n_g}, \quad g = 1, \ldots, G,$$

thereby generating $\bar{e}_{i,g}$ and $\bar{V}_{i,g}$. In our simulations, we set $\theta_2 = 0.7$.

For the parameters, we set $\beta = 0.3$, $\gamma = \tau = (1/\sqrt{d_w}) \times \iota_{d_w}$, where ι_{d_w} is a $d_w \times 1$ vector of ones. We specify geometrically decaying coefficients of IVs as $\pi = \left(\phi^0, \phi^1, \ldots, \phi^{K-1}\right)$, where K denotes the number of (many) base IVs and ϕ controls the relative weight assigned to each instrument. It is noted that $\phi = 0$ represents the case in which only the first instrument has identification strength, while $\phi = 1$ corresponds to the case in which each instrument has the same identification strength. We further normalize π to have $\|\pi\|_2 = \sqrt{\psi\sqrt{K}/n}$, where ψ controls the identification strength of the many IVs. The one-dimensional IV is constructed by taking the average of the many IVs; as ϕ approaches one, the identification strength of the low-dimensional IV becomes stronger since it is closer to the optimal instrument. We set the sample size at $n = 2{,}000$ and the number of clusters at $G = 500$, and then generate heterogeneous cluster sizes following a procedure similar to that in Djogbenou, MacKinnon, and Nielsen (2019). Specifically, for $g = 1, \cdots, G-1$, we set $n_g = \max\left\{1, n\exp(2g/G)/(1 + \sum_{g=1}^{G-1}\exp(2g/G))\right\}$ and then the size of the last cluster as $n_G = \max\left\{1, n - \sum_{g=1}^{G-1} n_g\right\}$. For the dimension of $\bar{Z}$, we consider $K = 100$ and $K = 500$, respectively. All the results below are based on $5,000$ simulations.

Figure 3 displays the power curves for our combination test ϕ_n^* along with those for the component Wald and jackknife LM tests, at different values of K (the dimension of the many IVs), ψ (which governs the identification strength of the many IVs), and ϕ (which controls the identification strength of the one-dimensional IV relative to the many IVs). We identified three main observations, each of which aligns with our large-sample theory. First, in every scenario, the combination test ϕ_n^* attains the correct size and is more powerful than each of the other two tests. In particular, as shown in Panel C, the power curve ϕ_n^* is never dominated by that of the Wald test, regardless of the strength of the many IVs, thereby underscoring the "free lunch" efficiency gains delivered by our combination test. Second, within Panels A and B of Figure 3, we observe that for fixed K and ψ, the power improvement of ϕ_n^* over the Wald test becomes more substantial as the identification strength of the one-dimensional IV weakens relative to that of the many IVs (i.e., as ϕ decreases). This is reflected in the widening gap between the power curves of ϕ_n^* and the Wald test. It emphasizes how many-IV-based LM and AR statistics contribute to the power enhancement. Third, in contrast to the cases in which the one-dimensional IV dominates many IVs in strength (first figure in Panel A or B) and the power curve of ϕ_n^* coincides with that of the Wald test, noticeable gaps persist between the power curves of ϕ_n^* and the LM test in the flipped cases (second and third figures in Panel A or B). These gaps highlight how the AR component contributes to power enhancement through its correlation with the LM statistic.

6 Conclusion

This paper proposes an inference approach that improves conventional estimation and inference in instrumental variables regressions using low-dimensional instruments (e.g., aggregated shift-share IVs) and their underlying high-dimensional base instruments. The procedure requires strong identification only for the low-dimensional IV regression while allowing the many-instrument specification to be weakly identified. We also provide a practical rule of thumb for when inference improves by at least 10%, based solely on the ratio of variances of the low- and high-dimensional IV estimators commonly reported in applications. Extensions to settings with many controls or diverging cluster sizes, alternative bootstrap procedures, and combinations with tests such as the sup-score test are left for future work.

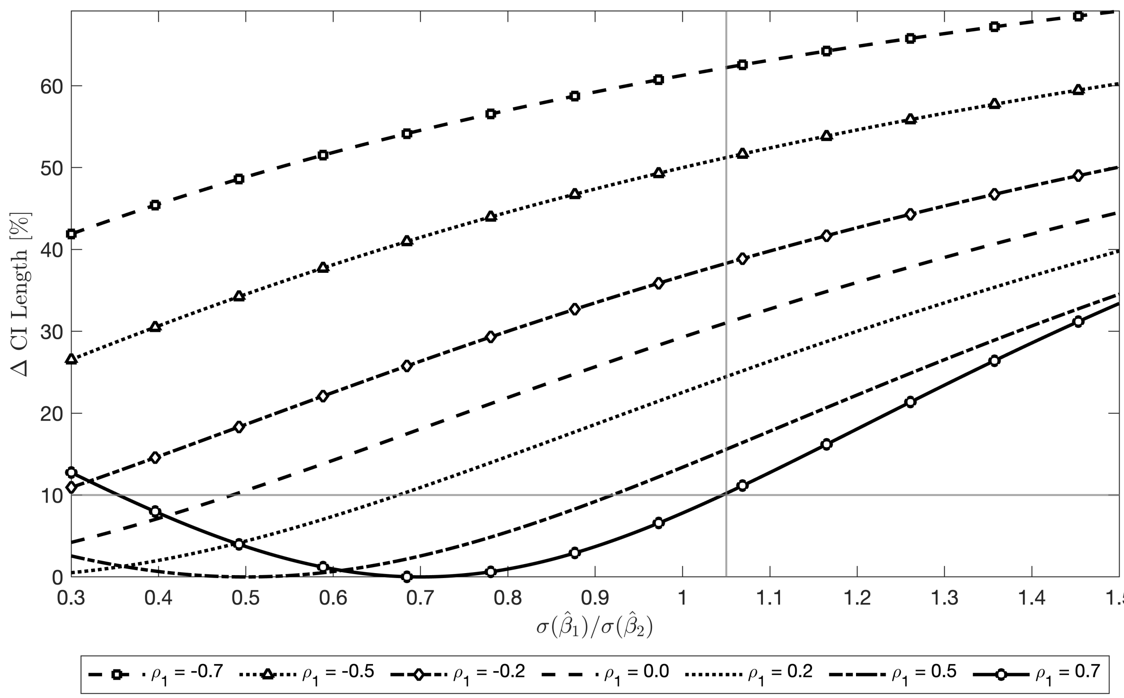

Figure 1: Theoretical lower bounds for percentage reduction in confidence interval length.

Notes: This figure plots the lower bound of the efficiency gain, given by (4.13), as a function of the standard deviation ratio, $\sigma(\hat\beta_1)/\sigma(\hat\beta_2)$, for various values of ρ_1, the limiting correlation between the Wald and (leave-one-cluster-out jacknife) LM statistics. The horizontal axis is the ratio of standard deviation of $\hat\beta_1$, the standard GMM estimator using low-dimensional IVs, and standard deviation of $\hat\beta_2$, the leave-one-cluster-out estimator using the many base IVs. The vertical axis is the reduction in the length of confidence interval in percentage points.

	College Equivalent Workers		High School Equivalent Workers	
	Yes	No	Yes	No
$\hat{\rho}_1$	0.588	0.446	0.408	0.576
$\hat{\rho}_2$	0.103	0.120	0.129	0.165
$\hat{\sigma}(\hat{\beta}_1)/\hat{\sigma}(\hat{\beta}_2)$	0.926	1.112	0.696	0.727
$\hat{\beta}_1$	-0.078	-0.080	-0.037	-0.024
Wald CI	(-0.103, -0.053)	(-0.107, -0.052)	(-0.051, -0.023)	(-0.037, -0.011)
$\hat{\beta}_2$	-0.066	-0.058	-0.043	-0.030
LM CI	(-0.093, -0.039)	(-0.083, -0.033)	(-0.063, -0.024)	(-0.048, -0.012)
$\hat{\beta}^*$	-0.072	-0.064	-0.039	-0.025
Comb. CI	(-0.095, -0.049)	(-0.087, -0.041)	(-0.052, -0.026)	(-0.038, -0.013)

Table 1: Point estimates and confidence intervals: Immigrant enclave.

Notes: This table reports the estimation and inference results for the immigrant enclave example using the Card (2009) dataset, shown separately for college equivalent workers and high school equivalent workers. Columns with "Yes" contain city-level controls, while columns with "No" do not. The point estimates are obtained from the standard two-stage least squares (TSLS) estimator with the one-dimensional Bartik instrument, $\hat{\beta}_1$, and, in addition, from the leave-one-cluster-out estimator, $\hat{\beta}_2$, which makes use of all base IVs. Wald CI and LM CI denote the confidence intervals based on $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The estimator $\hat{\beta}^*$ is the combined estimator for β, defined in Section 4.3. It is essentially the midpoint of the confidence interval in (4.11), which is obtained from our combination test and labeled as "Comb. CI" in the table. In addition, $\hat{\rho}_1$ and $\hat{\rho}_2$ denote estimates of the asymptotic correlation between the Wald and LM statistics, and between the LM and AR statistics, respectively. Finally, $\hat{\sigma}(\hat{\beta}_1)/\hat{\sigma}(\hat{\beta}_2)$ denotes the ratio of standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$. All displayed numbers are rounded to three decimal places.

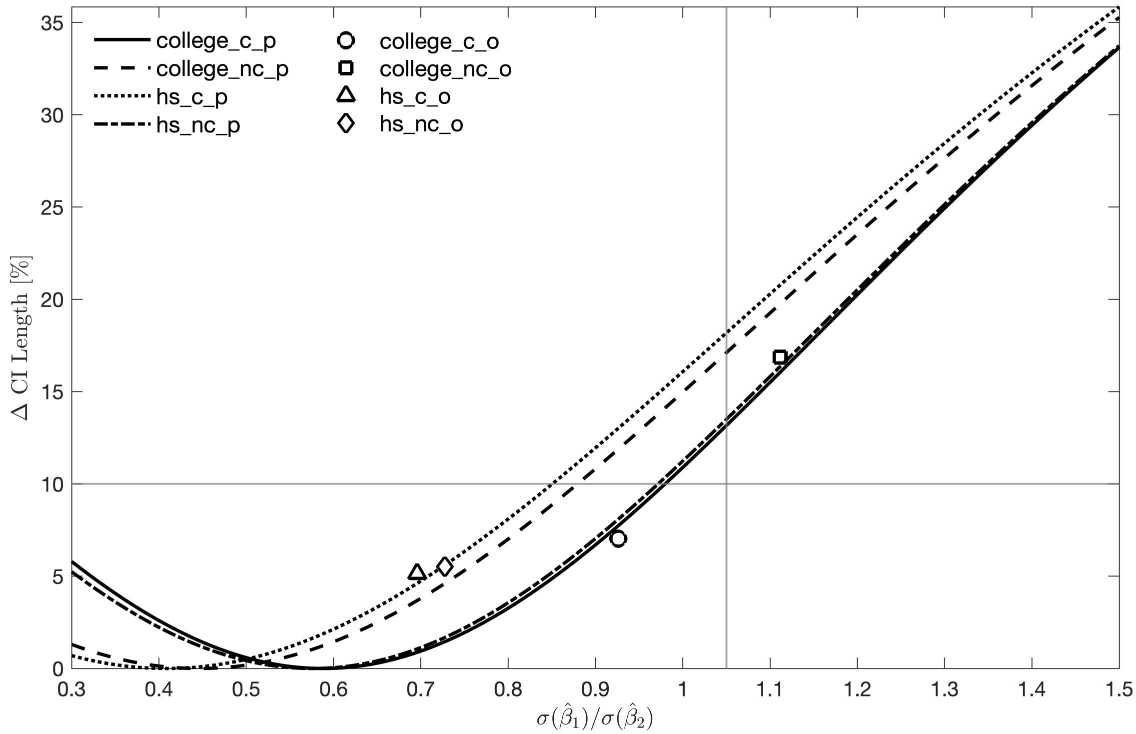

Figure 2: Realized percentage reduction in confidence interval length: Immigrant enclave.

Notes: This figure shows, for each specification in the immigrant enclave example, the observed percentage decrease in confidence interval length (Combined CI versus Wald CI, as in Table 1, and indicated by "o" in figure legends) plotted as a point against the standard error ratio ($\hat{\sigma}(\hat{\beta}_1)/\hat{\sigma}(\hat{\beta}_2)$ in Table 1). Also shown is the theoretical lower bound for the reduction (indicated by "p" in figure legends), analogous to Figure 1, but now computed using the specification-specific estimate $\hat{\rho}_1$, as reported in Table 1. Here, "college" refers to the specifications for college equivalent workers, and "hs" refers to the specifications for high school equivalent workers. "c" indicates that controls are included, whereas "nc" indicates that controls are not included. The horizontal axis is the ratio of standard deviations (errors) of $\hat{\beta}_1$ and $\hat{\beta}_2$. The vertical axis is the reduction in the length of confidence interval in percentage points. As a final remark, note that the actual numerical values of the relevant quantities in Table 1, rather than the rounded values shown there, are used to produce Figure 2.

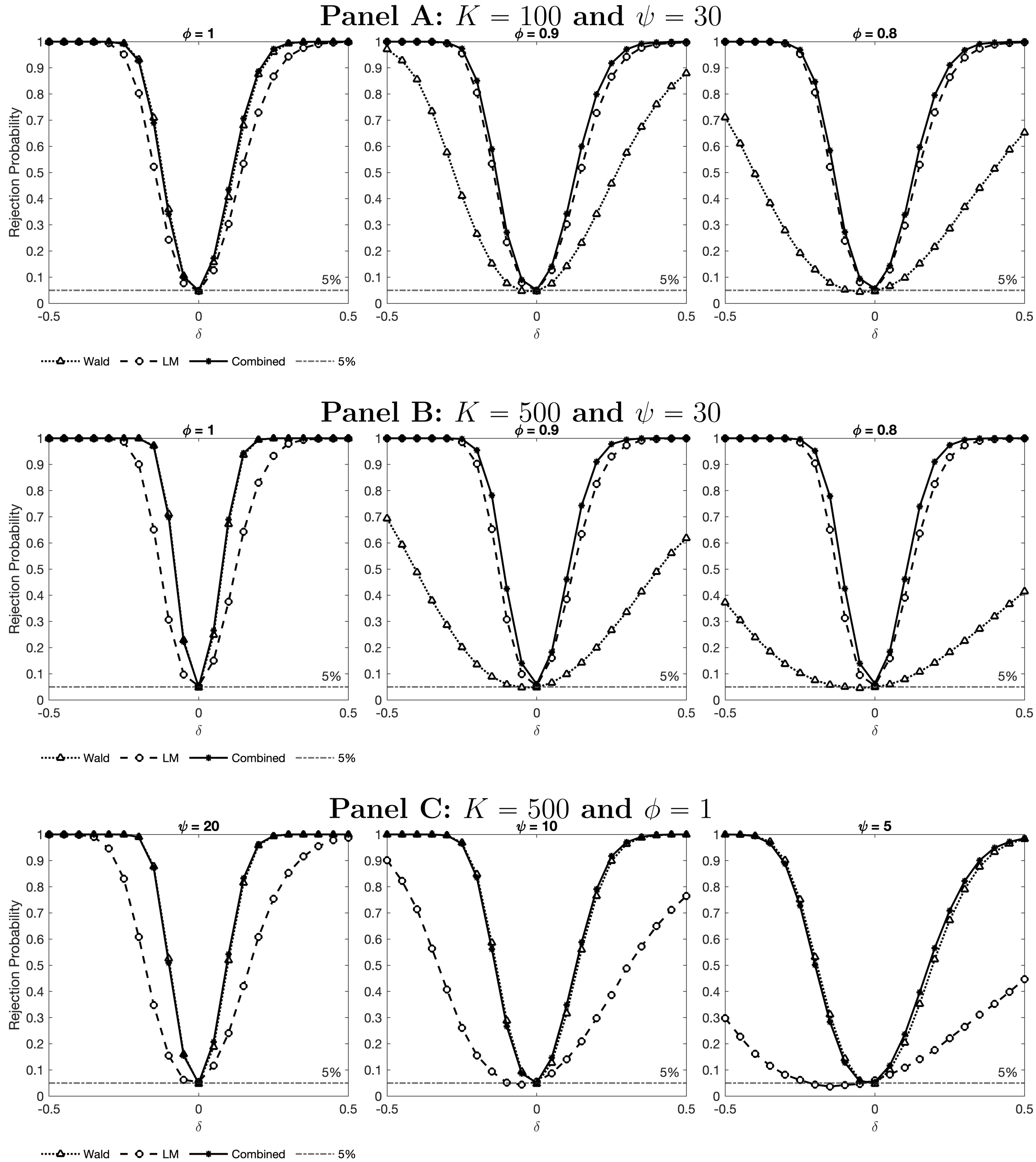

Figure 3: Power curves of the combination, Wald, and LM tests.

Notes: This figure displays the power curves for our combination test ϕ_n^* along with those for the component Wald and LM tests, at different values of K (the dimension of the many IVs), ψ (which governs the identification strength of the many IVs), and ϕ (which controls the identification strength of the one-dimensional IV relative to the many IVs). The horizontal axis represents the deviations in the parameter of interest from the maintained hypothesis, that is, we test $\mathcal{H}_0 : \beta = \beta_0$ against $\mathcal{H}_1 : \beta \neq \beta_0$, where $\delta = \beta - \beta_0$. See Section 5 for details of the simulation setup. All results are based on 5,000 simulations.

References

Adao, R., M. Kolesár, and E. Morales (2019): "Shift-share designs: Theory and inference," *The Quarterly Journal of Economics*, 134, 1949–2010.

Anatolyev, S. and N. Gospodinov (2011): "Specification testing in models with many instruments," *Econometric Theory*, 27, 427–441.

Anderson, T., N. Kunitomo, and Y. Matsushita (2010): "On the asymptotic optimality of the LIML estimator with possibly many instruments," *Journal of Econometrics*, 157, 191–204.

Andrews, D. W. (2017): "Identification-robust subvector inference," Tech. rep., Cowles Foundation Discussion Paper 2105.

Andrews, D. W. K. and J. H. Stock (2007): "Testing with many Weak Instruments," *Journal of Econometrics*, 138, 24–46.

Andrews, I. (2016): "Conditional linear combination tests for weakly identified models," *Econometrica*, 84, 2155–2182.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14, 57–67.

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earning?" *Quarterly Journal of Economics*, 106, 979–1014.

Bartik, T. J. (1991): *Who benefits from state and local economic development policies?*, Kalamazoo, MI: WE Upjohn Institute for Employment Research.

BEKKER, P. (1994): "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica*, 62, 657–681.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

BLANCHARD, O. J., L. F. KATZ, R. E. HALL, AND B. EICHENGREEN (1992): "Regional evolutions," *Brookings Papers on Economic Activity*, 1992, 1–75.

BOOT, T. AND D. NIBBERING (2024): "Inference on LATEs with Covariates," *arXiv preprint arXiv:2402.12607*.

BORUSYAK, K., P. HULL, AND X. JARAVEL (2022): "Quasi-experimental shift-share research designs," *The Review of economic studies*, 89, 181–213.

——— (2025): "A practical guide to shift-share instruments," *Journal of Economic Perspectives*, 39, 181–204.

CARD, D. (2009): "Immigration and Inequality," *American Economic Review: Papers & Proceedings*, 99, 1–21.

CARRASCO, M. (2012): "A regularization approach to the many instruments problem," *Journal of Econometrics*, 170, 383–398.

CARRASCO, M. AND G. TCHUENTE (2015): "Regularized LIML for many instruments," *Journal of Econometrics*, 186, 427–442.

CHAO, J. C. AND N. R. SWANSON (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

CHAO, J. C., N. R. SWANSON, J. A. HAUSMAN, W. K. NEWEY, AND T. WOUTERSEN (2012): "Asymptotic Distribution Of JIVE In A Heteroskedastic IV Regression With Many Instruments," *Econometric Theory*, 28, 42–86.

Chao, J. C., N. R. Swanson, and T. Woutersen (2023): "Jackknife estimation of a cluster-sample IV regression model with many weak instruments," *Journal of Econometrics*.

Djogbenou, A. A., J. G. MacKinnon, and M. O. Nielsen (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212, 393–412.

Donald, S. G. and W. K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161–1191.

Dovì, M.-S., A. B. Kock, and S. Mavroeidis (2024): "A Ridge-Regularized Jackknifed Anderson-Rubin Test," *Journal of Business & Economic Statistics*, 42, 1083–1094.

Dufour, J.-M. and M. Taamouti (2005): "Projection-based statistical inference in linear structural models with possibly weak instruments," *Econometrica*, 73, 1351–1365.

Evdokimov, K. S. and M. Kolesár (2018): "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects," *Working Paper*.

Frandsen, B., E. Leslie, and S. McIntyre (2025): "Cluster jackknife instrumental variables estimation," *The Review of Economics and Statistics*, forthcoming.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik instruments: What, when, why, and how," *American Economic Review*, 110, 2586–2624.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10, 487–526.

——— (2021): "A Powerful Subvector Anderson Rubin Test in Linear Instrumental Variables Regression with Conditional Heteroskedasticity," *arXiv preprint arXiv:2103.11371*.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression," *Econometrica*, 80, 2649–2666.

Han, C. and P. C. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74, 147–192.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398–422.

Hansen, C. and D. Kozbur (2014): "Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics*, 182, 290–308.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 3, 211–255.

Jiang, L., L. Li, K. Miao, and Y. Zhang (2025): "Adjustments with many regressors under covariate-adaptive randomizations," *Journal of Econometrics*, 249, 105991.

Kolesár, M. (2013): "Estimation in instrumental variables models with heterogeneous treatment effects," *Working Paper*.

——— (2018): "Minimum distance approach to inference with many instruments," *Journal of Econometrics*, 204, 86–100.

Kolesár, M., P. Min, W. Wang, and Y. Zhang (2026): "Cluster-Robust Inference for Quadratic Forms," *arXiv preprint arXiv:2602.13537*.

Kuersteiner, G. and R. Okui (2010): "Constructing optimal instruments by first-stage prediction averaging," *Econometrica*, 78, 697–718.

Kunitomo, N. (1980): "Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations," *Journal of the American Statistical Association*, 75, 693–700.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022): "Valid t-ratio Inference for IV," *American Economic Review*, 112, 3260–3290.

Lehmann, E. L. and J. P. Romano (2006): *Testing Statistical Hypotheses*, Springer Science & Business Media.

Ligtenberg, J. W. (2025): "Inference in IV models with clustered dependence, many instruments and weak identification," *arXiv preprint arXiv:2306.08559*.

Lim, D., W. Wang, and Y. Zhang (2024a): "A conditional linear combination test with many weak instruments," *Journal of Econometrics*, 238, 105602.

——— (2024b): "A Dimension-Agnostic Bootstrap Anderson-Rubin Test For Instrumental Variable Regressions," *arXiv preprint arXiv:2412.01603*.

Matsushita, Y. and T. Otsu (2024): "A jackknife Lagrange multiplier test with many weak instruments," *Econometric Theory*, 40, 447–470.

Mikusheva, A. and L. Sun (2022): "Inference with many weak instruments," *Review of Economic Studies*, 89, 2663–2686.

——— (2024): "Weak identification with many instruments," *The Econometrics Journal*, 27, C1–C28.

Morimune, K. (1983): "Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size," *Econometrica*, 51, 821–841.

Müller, U. K. (2011): "Efficient tests under a weak convergence assumption," *Econometrica*, 79, 395–435.

Navjeevan, M. (2024): "An Identification and Dimensionality Robust Test for Instrumental Variables Models," *arXiv preprint arXiv:2311.14892*.

Newey, W. K. and F. Windmeijer (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77, 687–719.

Okui, R. (2011): "Instrumental variable estimation in the presence of many moment conditions," *Journal of Econometrics*, 165, 70–86.

Sølvsten, M. (2020): "Robust estimation with many instruments," *Journal of Econometrics*, 214, 495–512.

Stock, J. and M. Yogo (2005): "Asymptotic distributions of instrumental variables statistics with many instruments," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 6, 109–120.

Wang, W. and F. Doko Tchatoka (2018): "On Bootstrap inconsistency and Bonferroni-based size-correction for the subset Anderson–Rubin test under conditional homoskedasticity," *Journal of Econometrics*, 207, 188–211.

Wang, W. and M. Kaffo (2016): "Bootstrap inference for instrumental variable models with many weak instruments," *Journal of Econometrics*, 192, 231–268.

Yap, L. (2025): "Inference with Many Weak Instruments and Heterogeneity," *arXiv preprint arXiv:2408.11193*.

Young, A. (2022): "Consistency without inference: Instrumental variables in practical application," *European Economic Review*, 147, 104112.