Quantification of Instruments' Strength*

Oriol González Marko Irisarri Santiago Iglesias Asier Beristain Manuel Cabado

Master Project†
2018/2019

Abstract

Weak identification is known to yield unreliable standard instrumental variables inference. A large literature has focused on addressing this issue by proposing methods to detect weak instruments, mainly through the first-stage F-statistic. This paper evaluates the weak identification in two leading empirical analyses by using the novel alternative approach developed by Ganics, Inoue and Rossi (2018), who base their tests on confidence intervals for the bias of the two-stage least squares estimator, and the size distortion of the associated Wald test. We illustrate the behavior of the tests in empirical settings, and compare how the conclusions differ to those using the standard tests. Our findings suggest that, in our empirical application, the results obtained using this approach are in line with those using previous tests in the literature, confirming it to be a robust alternative. An R package to directly compute these novel tests is also presented.

JEL Classification: C26, C52, C53, C88.

Keywords: Instrumental Variables; Weak Instruments; Concentration Parameter; Confidence Interval; Testing; R; Labor Supply; Economic Growth.

*We are thankful to Larbi Alaoui and David Puig for tutoring our work, as well as to all the participants in the seminar discussions of group B1 for their helpful comments and suggestions. We also appreciate Barbara Rossi's support. All errors are our own.

†Master's Degree in Economics, Barcelona Graduate School of Economics. Project group: B1.

Contents

1 Introduction

In this paper, we discuss the novel methodology proposed by Ganics, Inoue and Rossi (2018) for constructing confidence intervals to test for instruments' strength, namely for the bias of the two-stage least squares (2SLS) estimator, as well as for the size distortion of the associated Wald test, in various linear instrumental variables (IV) contexts. In order to do so, we introduce a new R package that retrieves such confidence intervals.

Instruments' strength is a common source of concern in IV estimations. It can be shown (see Stock et al., 2002) that when instruments are weakly correlated with endogenous regressors, the standard IV approach yields unreliable inference: point estimates are inconsistent and the size of the tests and confidence intervals are invalid.

A large literature has focused on addressing the weak identification problem by proposing methods to detect weak instruments. The most common approach is to use the first-stage F-statistic (or its generalization, for the case of multiple endogenous regressors), as proposed by Staiger and Stock (1997), Stock et al. (2002), and Stock and Yogo (2005). The procedure is simple: if the first-stage F-statistic is large enough, according to the appropriate critical values (see Stock and Yogo, 2005), then instruments are strong and standard inference is valid. However, a possible issue for applied researchers is that all these tests assume the errors to be homoskedastic and serially uncorrelated, which is not always the case in practice. Montiel-Olea and Pflueger (2013) provide an alternative which applies to general (heteroskedastic, serially correlated and/or clustered) errors, the so-called *effective* F-statistic, but only develop their work for the case of one endogenous regressor.

Ganics et al. (2018) propose a new approach to test the instruments' strength, based on constructing confidence intervals for the bias of the 2SLS estimator as well as for the size distortion of the associated Wald test. These novel tests present several advantages compared to those hitherto available. First, the methodology is robust to the presence of weak instruments. Second, and probably the most useful for applied researchers, the confidence intervals allow to answer *how* weak or strong instruments are, meaning that testing for weak instruments is not a binary decision anymore, as it is when using the first-stage F-statistic. Moreover, Ganics et al. (2018) point out that such confidence intervals "are straightforward and computationally easy to calculate, as they are obtained from inverting asymptotic chi-squared distributions". This simplicity contrasts with previous tests for weak instruments, which either involve computationally intensive bootstrapping (see Hansen, 1999), or whose distributions are typically asymptotically non-pivotal, as they depend on nuisance parameters that cannot be consistently estimated (see Staiger and Stock, 1997). Intuitively, this follows because previous tests are based on the difference between the estimate of the strength of identification and zero (no identification), thus containing information about the true strength of identification, which cannot be consistently estimated. The novel confidence intervals, instead, are based on the difference between the estimate and the true strength of identification, rather than the null hypothesis of no identification, and the limiting distribution of such difference does not depend on how weak instruments are.

Another important property addresses the need for applied researchers to find tests that are applicable to cases where errors are heteroskedastic or autocorrelated, as highlighted before by Chao et al. (2012) and Hausman et al. (2012). As previously mentioned, with the exception of Montiel-Olea and Pflueger (2013) and few others[1], tests for weak instruments in IV settings assume homoskedastic and serially uncorrelated errors. Montiel-Olea and Pflueger (2013) show that heteroskedasticity, serial correlation and/or clustering affect the weak instruments asymptotic distribution of the 2SLS estimator, what can further bias it and distort test sizes when instruments are weak, inducing the researcher to wrongly conclude that they are strong. The methodology by Ganics et al. (2018) can easily be applied when dealing with heteroskedastic and serially correlated disturbances; one will just replace the standard 2SLS estimator by the Heteroskedasticity and Autocorrelation Consistent (HAC) counterpart.

To the best of our knowledge, these recently proposed tests have not been subject to many empirical robustness checks, only the brief empirical analysis in Ganics et al.

[1] So far, there is no consensus on what tests should be used in over-identified and non-homoskedastic settings, and this literature is relatively recent. Very different alternatives have been proposed, what makes hard any comparison with the here presented tests. This is why we do not elaborate deeper on that. These alternatives include a variety of the Conditional Likelihood Ratio test for the non-homoskedastic case (see Kleibergen (2002), Andrews et al. (2004), Andrews (2016), Andrews and Guggenberger (2019), Andrews and Mikusheva (2016), among others). These have been proven to be efficient under strong instruments, however there is only simulation evidence on their power with weak instruments. Alternatively, some tests have been proposed that maximize the integral of the power function with respect to some weights (see, for instance, Moreira and Moreira (2015), Montiel-Olea (2017), or Moreira and Ridder (2017)). These latter raise the question of what the "right" weights are.

(2018) and the work by Ganics (2017), Chapter 3. Here, we elaborate on this by evaluating the empirical performance of this novel methodology, replicating previous studies dealing with potentially weak instruments and analysing how the conclusions about the instruments' strength compare to when using standard tests. We find very similar results regarding the conclusions of the instruments' strength as compared to when using previous tests in the literature, in our empirical applications. The paper is organized as follows. Section 2 presents the different tests proposed by Ganics et al. (2018), with special focus on the case of one endogenous regressor in the presence of homoskedasticity. Section 3 introduces the new package `girtest` in R that directly computes the confidence intervals for the bias and size distortion. Section 4 evaluates the empirical performance of such test. Section 5 concludes.

2 Test for Weak Instruments

Ganics et al. (2018) consider three different econometric frameworks, for which they propose the corresponding confidence intervals: the linear homoskedastic IV model, the heteroskedastic/autocorrelated linear IV model and the local projections-IV method. However, as the authors note, the latter is just a special case of an heteroskedastic and autocorrelated IV model.[2] Driven by the fact that most of the tests in the literature concern homoskedastic settings, and our aim is to analyse how the results compare when using the novel tests, we will only cover here the linear homoskedastic IV model, and leave the heteroskedastic/autocorrelated linear IV model for further work.

Throughout the paper, we will follow their notation, which is also commonly used in this literature. To this end, let T denote the sample size, and $\xrightarrow{p}$ and $\xrightarrow{d}$ convergence in probability and distribution, respectively. The vectorization operator is denoted by $\mathrm{vec}(\cdot)$ and $\otimes$ is the Kronecker product. The normal distribution with mean vector ψ and covariance matrix Ξ is denoted by $\mathcal{N}(\psi, \Xi)$, and $\chi^2_K(\tau)$ denotes a chi-squared distribution with K degrees of freedom and noncentrality parameter τ. For any $(T \times K)$ matrix A, $P_A \equiv A(A'A)^{-1}A'$ and $M_A \equiv I_K - P_A$, where I_K is the $(K \times K)$ identity matrix. We also adopt the convention that for a symmetric positive definite matrix B, $B = B^{1/2}B^{1/2}$ and $B^{-1} = B^{-1/2}B^{-1/2}$, where $B^{1/2}$ and $B^{-1/2}$ are the unique principal square roots.

[2]Subject, of course, to the validity of the structural vector moving average representation.

Following Staiger and Stock (1997) and Stock and Yogo (2005), consider the following two-stages least squares setup:

$$y = Y\beta + X\gamma + u \tag{1}$$

$$Y = Z\Pi + X\Phi + V \tag{2}$$

where y is a $(T \times 1)$ vector, Y is a $(T \times n)$ matrix of included endogenous variables, X is a $(T \times K_1)$ matrix of included exogenous variables (with one column of 1's if there is an intercept in (1)) and Z is a $(T \times K_2)$ matrix of excluded exogenous variables, i.e. instruments. β is an $(n \times 1)$, while γ is a $(K_1 \times 1)$ vector of coefficients. Π is a matrix of coefficients of dimension $(K_2 \times n)$, and Φ is a $(K_1 \times n)$ matrix of coefficients. The vector of disturbances u is $(T \times 1)$, while V is a $(T \times n)$ matrix of errors.

It is convenient to define $X_t = (X_{1t}, \dots, X_{K_1 t})'$, $Z_t = (Z_{1t}, \dots, Z_{K_2 t})'$, $V_t = (V_{1t}, \dots, V_{nt})'$ and $\underline{Z}_t = (X_t', Z_t')'$ as the vectors of the t-th observations of the respective variables for $t = 1, \dots, T$. Also, let Σ and Q denote the population second moment matrices, where

$$\Sigma = \mathbb{E}\left[\begin{pmatrix} u_t \\ V_t \end{pmatrix} \begin{pmatrix} u_t & V_t' \end{pmatrix}\right] = \begin{bmatrix} \sigma_{uu} & \Sigma_{uV} \\ \Sigma_{Vu} & \Sigma_{VV} \end{bmatrix} \tag{3}$$

$$Q = \mathbb{E}[\underline{Z}_t \underline{Z}_t'] = \begin{bmatrix} Q_{XX} & Q_{XZ} \\ Q_{ZX} & Q_{ZZ} \end{bmatrix} \tag{4}$$

To develop the weak instruments asymptotics, following Staiger and Stock (1997), the authors model Π as local to zero[3] (with a Pitman drift) and put certain restrictions on the moments, namely the required homoskedasticity and serial uncorrelation of the errors (part (c) of Assumption M). Formally:

Assumption $\mathbf{L}_{\Pi}$: $\Pi = \Pi_T = C/\sqrt{T}$, where C is a fixed $K_2 \times n$ matrix.

Assumption M: The following limits hold jointly for fixed K_2 as $T \to \infty$:

(a) $(T^{-1}u'u, T^{-1}V'u, T^{-1}V'V) \xrightarrow{p} (\sigma_{uu}, \Sigma_{Vu}, \Sigma_{VV})$;

(b) $T^{-1}\underline{Z}'\underline{Z} \xrightarrow{p} Q$, where Q is positive definite;

(c) $(T^{-1/2}X'u, T^{-1/2}Z'u, T^{-1/2}X'V, T^{-1/2}Z'V) \xrightarrow{d} (\Psi_{Xu}, \Psi_{Zu}, \Psi_{XV}, \Psi_{ZV})$, where $\Psi \equiv [\Psi_{Xu}', \Psi_{Zu}', \mathrm{vec}(\Psi_{XV})', \mathrm{vec}(\Psi_{ZV})']' \sim \mathcal{N}(0, \Sigma \otimes Q)$, where Σ is positive definite.

[3] This implies that the first-stage F-statistic, F, is $O_p(1)$, what would explain why the mean of F testing $\Pi = 0$ in equation (2) is small or moderate even if T is large.

To simplify exposition, the exogenous regressors X are projected out, by Frisch-Waugh Theorem. Let $Y^{\perp} \equiv M_X Y$, $Z^{\perp} \equiv M_X Z$ and $V^{\perp} \equiv M_X V$. Also, let $V_t^{\perp}$ be the transpose of the t-th row of $V^{\perp}$, and similarly for $Z_t^{\perp}$. The first-stage equation in (2) rewrites as

$$Y^{\perp} = Z^{\perp}\Pi + V^{\perp} \tag{5}$$

Furthermore, define $\Omega \equiv Q_{ZZ} - Q_{ZX}Q_{XX}^{-1}Q_{XZ} = Q_{Z^{\perp}Z^{\perp}}$, where $Q_{Z^{\perp}Z^{\perp}} \equiv \mathbb{E}[Z_t^{\perp}Z_t^{\perp\prime}]$ and $\hat{\Omega} \equiv Z^{\perp\prime}Z^{\perp}/T$. Finally, let $\hat{\Pi}_T = (Z^{\perp\prime}Z^{\perp})^{-1}Z^{\perp\prime}Y^{\perp}$ denote the OLS estimator of Π in equation (5).

Ganics et al. (2018) develop their confidence intervals for the strength of identification based on the concentration parameter,[4] which is defined as

$$\mu_{t,K_2}^2 \equiv \frac{1}{K_2}\Pi' Z^{\perp\prime} Z^{\perp}\Pi/\sigma_{VV} \xrightarrow{p} \frac{1}{K_2}C'\Omega C/\sigma_{VV} \equiv \mu_{K_2}^2 \tag{6}$$

As Stock et al. (2002) note, a useful interpretation of $\mu_{K_2}^2$ is in terms of the first-stage F-statistic, F. If we let $\tilde{F}$ denote the infeasible counterpart of F (computed using the true σ_{VV}), then $K_2\tilde{F} \xrightarrow{d} \chi_{K_2}^2(K_2\mu_{K_2}^2)$ and $\mathbb{E}[\tilde{F}] = \mu_{K_2}^2 + 1$. Therefore, when T grows, F and $\tilde{F}$ become increasingly closer. In that case $\mathbb{E}[F] \cong \mu_{K_2}^2 + 1$, so that $F - 1$ can be thought of as an estimator for $\mu_{K_2}^2$.

Ganics et al. (2018) provide two different approaches to obtain the confidence intervals for the bias and size distortion: the non-central χ^2 and the projection method. The former is only applicable in cases where there is only one endogenous regressor, while the latter is applicable in general cases with potentially multiple endogenous regressors. However, they show by means of Monte Carlo simulations that in general the projection method yields more conservative results, in the sense that the coverage rates are not as close to the nominal levels as for the non-central χ^2. For this reason, we will focus on the non-central χ^2 in the empirical analysis, and hence provide here the full derivation for it, whereas only the intuition of the methodology for the projection method.

2.1 The Case of One Endogenous Regressor

The non-central χ^2 method is essentially based on the asymptotic distribution of the OLS estimator of Π in equation (5), $\hat{\Pi}_T$. Under Assumptions L_{Π} and M, its limiting

[4] Or the minimum eigenvalue of the concentration matrix, when there are multiple endogenous regressors. See Cragg and Donald (1993) and Stock and Yogo (2005) for more on that.

distribution is given by

$$\sqrt{T}\hat{\Pi}_T \xrightarrow{d} \mathcal{N}(C, \sigma_{VV}\Omega^{-1}) \tag{7}$$

which by Slutsky's theorem implies that

$$m_T \equiv \hat{\Omega}^{1/2}\hat{\sigma}_{VV}^{-1/2}\sqrt{T}\hat{\Pi}_T \xrightarrow{d} \mathcal{N}(\Omega^{1/2}C\sigma_{VV}^{-1/2}, I_{K_2}) \tag{8}$$

yielding that the statistic $f_T \equiv m_T' m_T$ is asymptotically distributed as

$$f_T \xrightarrow{d} \chi^2_{K_2}(K_2\mu^2_{K_2}) \tag{9}$$

Now, because the chosen parameter for the strength of identification is the concentration parameter, and K_2 is known, we simply need to obtain a confidence interval for the noncentrality parameter of the χ^2 distribution in equation (9).

In order to do so, one can simply follow Kent and Hainsworth (1995). In particular, as the authors recommend, Ganics et al. (2018) use the 'symmetric range' method to construct such confidence interval. The full derivation can be found in Appendix A of Ganics et al. (2018). Applying this method we obtain $\mathrm{CI}^{\mu^2_{K_2}}_{1-\alpha} \equiv [l^{\mu^2_{K_2}}_{1-\alpha}, u^{\mu^2_{K_2}}_{1-\alpha}]$, which is a $(1-\alpha)$ level asymptotic confidence interval for $\mu^2_{K_2}$. Also, define

$$l^b_{1-\alpha} \equiv b(u^{\mu^2_{K_2}}_{1-\alpha}; n, K_2) \qquad u^b_{1-\alpha} \equiv b(l^{\mu^2_{K_2}}_{1-\alpha}; n, K_2) \tag{10}$$

$$l^s_{1-\alpha} \equiv s(u^{\mu^2_{K_2}}_{1-\alpha}; n, K_2) \qquad u^s_{1-\alpha} \equiv s(l^{\mu^2_{K_2}}_{1-\alpha}; n, K_2) \tag{11}$$

where b$(\cdot;\cdot)$ and s$(\cdot;\cdot)$ denote the bias and the size distortion of the 2SLS estimator, respectively.[5] These constitute the endpoints of the $(1-\alpha)$ level asymptotic confidence intervals for the bias (equation (10)) and size distortion (equation (11)).

The main result is summarized in Proposition 1 of Ganics et al. (2018), which we proceed to reproduce. Under Assumptions L_Π and M, $\mathrm{CI}^{\mu^2_{K_2}}_{1-\alpha}$ is an asymptotically valid $(1-\alpha)$ level confidence interval for $\mu^2_{K_2}$, that is

$$\lim_{T\to\infty} \mathbb{P}\left(\mu^2_{K_2} \in \mathrm{CI}^{\mu^2_{K_2}}_{1-\alpha}\right) = 1-\alpha \tag{12}$$

Furthermore, $[l^b_{1-\alpha}, u^b_{1-\alpha}]$ and $[l^s_{1-\alpha}, u^s_{1-\alpha}]$ are $(1-\alpha)$ level asymptotic confidence intervals

[5] Note that the bias and size distortion are decreasing functions, as explained below.

for the bias and size distortion, respectively. Formally:

$$\lim_{T\to\infty} \mathbb{P}\Big(b(\mu^2_{K_2}; n, K_2) \in [l^b_{1-\alpha}, u^b_{1-\alpha}] \Big) = 1 - \alpha \tag{13}$$

$$\lim_{T\to\infty} \mathbb{P}\Big(s(\mu^2_{K_2}; n, K_2) \in [l^s_{1-\alpha}, u^s_{1-\alpha}] \Big) \geq 1 - \alpha \tag{14}$$

The equality in equation (13) follows by the result in Theorem B2 in Skeels and Windmeijer (2016), who prove that in the case we are considering of one endogenous regressor, the bias $b(\mu^2_{K_2}; 1, K_2)$ is a *strictly* decreasing continuous function of $\mu^2_{K_2}$. Moreover, the simulations in Stock and Yogo (2005) suggest that the size $s(\mu^2_{K_2}; n, K_2)$ is also strictly decreasing, what means that the weak inequality in (14) will turn into an equality, and thus the proposed asymptotic confidence interval will not be conservative.

2.2 The Case of Potentially Multiple Endogenous Regressors

The projection method is based on the projection argument developed by Dufour (1990, 1997). The intuition behind this approach is that given a confidence set $C_\mu(\alpha)$ with level $(1-\alpha)$ for the parameter vector μ, one can obtain confidence sets for general transformations g in $\mathbb{R}^m$ of this vector. Since $s \in S \Rightarrow g(s) \in g(S)$ for any set S, we have

$$\mathbb{P}\big(\mu \in C_\mu(\alpha)\big) \geq 1 - \alpha \quad \Longrightarrow \quad \mathbb{P}\big(g(\mu) \in g(C_\mu(\alpha))\big) \geq 1 - \alpha \tag{15}$$

where $g(C_\mu(\alpha)) = \{s \in \mathbb{R}^m : \exists \mu \in C_\mu(\alpha), g(\mu) = s\}$. Hence $g(C_\mu(\alpha))$ is a conservative confidence set for $g(\mu)$ with level $(1-\alpha)$. However, computing $g(C_\mu(\alpha))$ may in general be a costly numerical exercise (see, for instance, Dufour and Jasiak (2001)).

Ganics et al. (2018) use this approach by exploiting the mapping from the parameter summarizing the strength of identification to the bias and size distortion. In particular, the parameter for the strength of identification they use in this case is the minimum eigenvalue of the concentration matrix, of which, as Stock and Yogo (2005) show, the worst-case asymptotic bias relative to the OLS estimator and worst-case asymptotic size distortion of the associated Wald test – where the worst-case corresponds to the maximum of these quantities over all possible degrees of simultaneity between the error terms in equations (1) and (2) – are continuous and decreasing functions, albeit no closed-form expression is known for a general number of endogenous regressors.

3 Package `girtest` in R

One of the main motivations of this project is not only to illustrate why we think this new approach is interesting, but also to facilitate other researchers its implementation in their own projects. We have developed a statistical package in the statistical computing software R that directly computes the above derived test (section 2.1). This section introduces the package.

The R package `girtest` contains the function `girtest`. This function allows to retrieve three different 95% asymptotic level confidence intervals: one for the concentration parameter; one for the bias of the 2SLS estimator; and one for the size distortion of the associated Wald test. It takes two arguments: the F-statistic of the first-stage regression on the strength of identification, `Fstat`, and the number of instruments (excluded exogenous variables) used, `K_2`. The critical values for the bias and size distortion are obtained from Appendix D in Ganics et al. (2018), Table D.7 and Table D.10, respectively. It must be noted that such function is only applicable in homoskedastic settings where there is only one endogenous regressor.

The first time the function is to be used, the package needs to be installed and loaded. This procedure slightly differs from the standard one because the package is (for now) uploaded on GitHub, instead of CRAN, since the latter requires a long procedure of official validation. The process is still very simple though. First install and load the package `devtools`, by typing `install.packages("devtools")` followed by `library(devtools)`. Now we are ready to install our package `girtest`. To install it, we need to call the package stored in the GitHub repository "girtest" of the user OriolGC. Run the command `install_github("OriolGC/girtest")` to install the package and load it with `library(girtest)`.[6] Now the function is ready to be used.

The utilisation procedure is extremely simple. First, the researcher estimates the first-stage regression and obtains the standard F-statistic. Then, knowing how many instruments are being used, she fills in the arguments of the function with that information. Here we present a very simple example. Suppose the researcher obtains an F-statistic of 14.6 in the first-stage regression, when using 4 instruments. Then, she would type

[6] As usual, the `library(girtest)` command is required every time R is opened.

`girtest(Fstat = 14.6, K_2 = 4)` in the R command window.[7] In that case, the following output is printed:[8]

```
> girtest(Fstat = 14.6, K_2 = 4)

Confidence intervals by Ganics, Inoue and Rossi (2018)

Concentration Parameter [ 8.01 ; 23.09 ]
Bias                    [ 0.03 ; 0.06 ]
Size distortion         [ 0.04 ; 0.1 ]
```

An important aspect we have taken into account is that not all applied researchers tend to use R, what means that they probably estimate the first-stage regression in another programming language. For this reason, we have decided the `girtest` function not to be a direct postestimation command (taking directly the F-statistic from the command `ivreg` in R, or the number of instruments) but we have rather allowed that any researcher were able to introduce their estimated parameters, maybe from other computing programs. We believe this makes the implementation of the test even easier, and makes sure that the lack of knowledge in R programming is not a restriction for its computation.

3.1 Performance Test of the `girtest` Function

Before moving on to the empirical section, we first demonstrate the good performance of the R function we have constructed. To do so, we replicate Table 6 in Ganics et al. (2018). The results are shown in Figure I.

	IES ψ	IES ψ^{-1}
TSLS estimate (standard error)	0.06 (0.086)	0.68 (0.474)
95% Confidence Interval for bias	[0.021, 0.058]	[0.069, 0.786]
95% Confidence Interval for size distortion	[0.034, 0.090]	[0.105, 0.816]
F-statistic	15.53	2.93
Critical value (5% bias)	16.85	16.85
Critical value (10% bias)	**10.27**	10.27
Critical value (5% size distortion)	24.58	24.58
Critical value (10% size distortion)	**13.96**	13.96

(a) Table 6 in Ganics et al. (2018)

```
> girtest(Fstat = 15.53, K_2 = 4)

Confidence intervals by Ganics, Inoue and Rossi (2018)

Concentration Parameter [ 8.7 ; 24.25 ]
Bias                    [ 0.03 ; 0.06 ]
Size distortion         [ 0.03 ; 0.09 ]
> girtest(Fstat = 2.93, K_2 = 4)

Confidence intervals by Ganics, Inoue and Rossi (2018)

Concentration Parameter [ 0.25 ; 7.31 ]
Bias                    [ 0.07 ; 0.79 ]
Size distortion         [ 0.11 ; 0.9 ]
```

(b) Replication using the function `girtest`

Figure I: Assessing the performance of the function `girtest`

From Table 6 in Ganics et al. (2018), we only need to focus on Panel B, which provides the estimates using the non-central χ^2 approach. We can see that the results using the

[7] The name of the arguments, '`Fstat =`' and '`K_2 =`', can indeed be ignored, but we include them here to make the exposition clearer.

[8] For how to interpret the results, see sections 4.1 and 4.2 or Ganics et al. (2018).

`girtest` function are very accurate up to the second decimal, which is the information provided in Appendix D in Ganics et al. (2018) regarding the critical values. With this result, we ensure the robustness of the analysis in the following section.

Last, mention that we have decided to provide a confidence interval for the concentration parameter as well. We believe this is of important relevance because $\mu^2_{K_2}$ can be shown to play the role of the sample size (see Rothenberg, 1984). In order words, $\mu^2_{K_2}$ is a nuisance parameter which measures the amount of information the data have about the parameter of interest β in equation (1). Hence if $K_2\mu^2_{K_2}$ is large, $\sqrt{K_2\mu^2_{K_2}}(\hat{\beta}_{2SLS} - \beta)$ will be approximately normal, whereas if $K_2\mu^2_{K_2}$ is small, then the distribution is non-standard, what comes across as crucial for inference.

4 Empirical Evaluation

In this section, we use the test from Ganics et al. (2018) described above to evaluate the robustness of the results in two important empirical analyses. The first is Levine et al. (2000) and the second is Angrist and Evans (1998). To do so, we replicate the two papers in the two-stage least squares literature, and then we apply the proposed `girtest` function to assess the instruments' strength, and analyse how our conclusions differ from those of the original authors.

4.1 Financial Intermediation and Growth

Determining whether the relationship between the role of financial intermediaries and economic growth is mere correlation or can be interpreted as causal has been a widely studied topic in previous literature. One of the main arguments to defend the economic importance of financial institutions is their key role on offering products and contracts that allow individuals to insure themselves against information asymmetries (see Boyd and Prescott, 1986), therefore leading to a better allocation of resources. However, other studies show that this better allocation might lower saving rates enough to cancel out the positive effect. This important division in previous literature results has motivated researchers to keep studying this topic. Identification of a causal effect could be key for policy makers when modifying the current regulation. In this section we consider the paper by Levine et al. (2000), which intends on evaluating whether the exogenous component of financial intermediary development has an influence on economic growth.

They try to correct for the endogeneity issues arising in the financial development measures via an IV approach, using as instruments legal origin indicators.

Levine et al. (2000) defend the validity of such approach because of the following reasons. The exogeneity of the legal origin indicators, identifying English, German, French or Scandinavian systems, is justified by the fact that these were primarily spread through conquest and imperialism. The arguments in favor of the relevance condition are based on the findings in Porta et al. (1998). Namely, they claim that differences in the legal origin lead to differences in the legal rules covering secured creditors, the efficiency of contract enforcement, and the quality of accounting standards, what characterize financial intermediary activities.

Following their identification strategy,[9] we consider the 2SLS approach where in the first stage the legal origin of each country is used as an instrument for the indicators of financial intermediary development, controlling also for the level of income. Four possible legal origins are considered due to their worldwide importance: English, Napoleonic, German and Scandinavian. In the second stage, real per capita GDP growth is regressed on two different measures of financial intermediary development, Liquid Liabilities (LL) and Private Credit (PC), one at a time, together with a set of controls conditioning the information set. In particular, the model takes the following form:

$$y_i = \alpha + \beta E_i + \gamma' \mathbf{X}_i + \varepsilon_i \qquad \text{(2nd stage)}$$

$$E_i = \pi_0 Scand_i + \pi_1 Ger_i + \pi_2 Eng_i + \pi_3 Fr_i + \delta Inc_i + \eta_i \qquad \text{(1st stage)}$$

where y_i is the real per capita GDP growth rate of country i. The endogenous measures of financial development are $E_i = \{PC_i, LL_i\}$, where PC_i is the credit by deposit money banks and other financial institutions to the private sector divided by GDP, times 100, and LL_i is the liquid liabilities of the financial system (currency plus demand and interest-bearing liabilities of banks and nonbank financial intermediaries) divided by GDP, times 100. The vector of controls, based on the "simple conditioning information set", is represented by $\mathbf{X}_i$. All the regressors in the first-stage regression are dummy variables indicating the legal origin of the country, except for Inc_i, which is the logarithm of real per capita GDP in 1960. Namely, $Scand_i$, Ger_i, Eng_i and Fr_i indicate Scandinavian, German, English and Napoleonic legal origin, respectively. The

[9] Although they estimate the model via GMM, they explicitly indicate in their footnote 10 that "the 2SLS procedure produces the same conclusions".

parameters ε_i and η_i are the error terms.

Table I: Financial Intermediation and Growth

	Private Credit	Liquid Liabilities
2SLS estimate (standard error)	2.52 (0.81)	1.72 (0.84)
95% Confidence interval for bias	[0.03 ; 0.24]	[0.03 ; 0.24]
95% Confidence interval for size distortion	[0.05 ; 0.31]	[0.05 ; 0.27]
F-statistic	5.85	6.14
Critical value (20% bias)	6.46	6.46
Critical value (30% bias)	**5.39**	**5.39**
Critical value (20% size distortion)	9.54	9.54
Critical value (25% size distortion)	7.80	7.80

Note: Following the notation in section 2, here we have $n = 1$ and $K_2 = 3$. The upper panel reports the confidence intervals for the bias and size distortion in the Levine et al. (2000) economic growth regressions. The lower panel displays the first-stage F-statistic and the corresponding critical values (at the 5% significance level) for the bias and size distortion (nominal level for the Wald test is 5%) following Stock and Yogo (2005). Critical values in bold correspond to strong instruments. The 2SLS estimators are computed under the "simple conditioning information set" considered in Levine et al. (2000).

The estimation results are displayed in table I. The upper panel shows the 95% level confidence intervals for the bias of the 2SLS estimator and the size distortion of the associated Wald test. Under either endogenous variable, the results suggest a bias of the 2SLS estimator between 3% and 24%. At the same time, we estimate a size distortion between 5% and 31%, when using Private Credit, whereas a size distortion between 5% and 27% when using Liquid Liabilities. Applied researchers might not feel comfortable dealing with such numbers; for instance, realizing that your test has a 31% larger size than advertised. We cannot therefore conclude that the instruments for the legal origin of a given country are strong under any of the two specifications in order to explain economic growth, meaning that inference needs to be taken with caution.

The lower panel of the table reports the results for the binary testing procedure, based on the critical values of the first-stage F-statistic by Stock and Yogo (2005). We see that there is no evidence to reject the null at the 5% significance level that asymptotically the bias of the 2SLS estimator is at most 20% of the bias of the OLS in the worst case. However, the null at the 5% level that the bias of the 2SLS estimator is at most 30% of the bias of the OLS in worst case is rejected. Likewise, there is no evidence to reject the null that when performing a Wald test at the 5% nominal level, asymptotically the test would have a 20% or 25% larger size than claimed in the worst case. The results from this approach coincide with those of the confidence intervals.

Given the analysis using both procedures, the researcher would have well-founded evidence to suspect that the instruments are indeed weak. This contrasts the conclusions in the original paper, where they claimed that the exogenous component of financial intermediary development is positively associated with economic growth, and this "is not due to potential biases induced by omitted variables, simultaneity or reverse causation". Even the effect on economic growth of "their preferred indicator" of financial intermediary development, Private Credit, fails to be consistently estimated under this setting because of weak identification.

4.2 Children and Their Parents' Labor Supply

Research on the labor-supply consequences of childbearing is important for a number of theoretical and practical reasons, but complicated because of the endogeneity of fertility. Different approaches have been proposed to deal with such issue, but none has managed to steer clear of the skepticism regarding the causal interpretation of the results. Here we consider the analysis by Angrist and Evans (1998). They mainly contribute to the literature on fertility and labor-supply by proposing a new instrumental variables strategy based on the sex mix to resolve this endogeneity problem.

Angrist and Evans (1998) aim at analyzing the effects of childbearing on (mainly married) women's labor-supply outcomes. To that end, they analyze the effect of having more than two children on the outcomes hours worked, hours worked per week, labor income and whether the parent worked during that year. To account for the endogeneity of having more than two children, the sex composition of the first two children is used as instrument(s). They claim this instrument approach is valid: the exogeneity condition is defended by arguing that the sex of the child is a random assignment. As for the relevance condition, what we study here, they point to the growing evidence that families with children of the same sex are more likely to have a third one.

The estimation strategy of the paper is 2SLS. In the first stage, they control for the effect of the sex mix on the decision of having more than two children, in order to consistently estimate the effect of having more than two children on labor-supply variables of the parents, in the second stage. In particular, the model we consider for replication is the following:

$$y_i = \alpha_0' \mathbf{w}_i + \pi_1 s_{1i} + \beta_1 x_i + \varepsilon_i \qquad \text{(2nd stage)}$$

$$x_i = \pi_0' \mathbf{w}_i + \pi_2 s_{1i} + \gamma_0 (Twoboys_i) + \gamma_1 (Twogirls_i) + \eta_i \qquad \text{(1st stage)}$$

where y_i is the number of weeks worked, x_i is the endogenous regressor indicating whether the woman has more than two kids, s_{ji} is a dummy for the jth child being a boy, such that $Twoboys_i = s_{1i}s_{2i}$ and $Twogirls_i = (1 - s_{1i})(1 - s_{2i})$, $\mathbf{w}_i$ is a vector of exogenous regressors, and η_i and ε_i are the error terms.

One may have doubts whether is it true that families with children of the same sex are more likely to have a third, and hence suspect that the instruments may not be strong enough. Computing the first-stage F-statistic for the null hypothesis of no identification, we find a value of $F = 715.13$.[10] The confidence intervals for the bias and size distortion of the 2SLS estimator and the associated Wald test in this model, computed using the `girtest` function, are shown in table II.

Table II: Children and Their Parents' Labor Supply

2SLS estimate (standard error)	-5.16 (1.20)
95% Confidence interval for bias	[0.00; 0.00]
95% Confidence interval for size distortion	[0.00; 0.00]
F-statistic	715.13
Critical value (5% bias)	**9.02**
Critical value (5% size distortion)	**19.93**

Note: Following the notation in section 2, here we have $n = 1$ and $K_2 = 2$. The upper panel reports the confidence intervals for the bias and size distortion in the Angrist and Evans (1998) parent's labor supply regression; Table 6, column 6. The lower panel displays the first-stage F-statistic and the corresponding critical values (at the 5% significance level) for the bias and size distortion (nominal level for the Wald test is 5%) following Skeels and Windmeijer (2016) and Stock and Yogo (2005). Critical values in bold correspond to strong instruments.

The lower panel of the table suggests that the proposed instruments are indeed strong according to the binary testing procedure. Based on the critical values for the F-statistic provided by Stock and Yogo (2005) and Skeels and Windmeijer (2016), we can reject the null at the 5% significance level that asymptotically the bias of the 2SLS estimator

[10]Note that the original paper did not include the first-stage F-statistic, although the t-statistics on individual parameters hinted at a strong significance.

is at most 5% of the bias of the OLS in the worst case. Similarly, we can reject the null that when performing a Wald test at the 5% nominal level, asymptotically the test would have a 5% larger size than claimed in the worst case. The upper panel displays the 95% confidence intervals for the bias and size distortion of the 2SLS estimator as proposed by Ganics et al. (2018). The results are in line with the previous findings. Note that given the almost zero-measure of the confidence intervals, we can claim that the bias of the 2SLS estimator and size distortion of the associated Wald test are both virtually non-existent.

The main conclusion from our analysis is that the researcher should in this case be confident that the model does not suffer from weak identification and hence the standard IV inference is reliable. Namely, under the assumptions of the original work, the instruments $Twoboys$ and $Twogirls$ satisfy the relevance condition needed to be considered valid instruments for the fertility indicator of having more than two children.

To finish with the discussion of this empirical case, we believe important to point out the sentence "under the assumptions of the original work" in the last paragraph. In this analysis, we did not get into the detail of whether the entire model is the right one to study the labor supply of the parents, but instead focused on the validity of the instruments used. Indeed, some readers may be skeptical about having such a large first-stage F-statistic, despite the large number of observations. The work in Angrist and Evans (1998) relies on a very strong assumption, homoskedasticity, which they never test or provide any justification for why to believe it holds. As mentioned in section 1, Montiel-Olea and Pflueger (2013) and Young (2019) stress the fact that falsely assuming a specific form of the error variance can mislead the researcher to think she is dealing with strong instruments. Indeed, different tests for heteroskedasticity, such as the ones proposed by Breusch and Pagan (1979), White et al. (1980) or Cook and Weisberg (1983), which were available at that time, lead to reject the null hypothesis of homoskedasticity in this sample, warning us about the reliability of the conclusions in the original paper. The results are provided in the appendix.

5 Conclusions

In this paper we evaluate the performance of the novel test proposed by Ganics et al. (2018) to assess the instruments' strength in various linear homoskedastic IV contexts, and present an R function to directly compute it. The contribution of this paper is mainly twofold: first, it should serve to evaluate the robustness of the empirical results in two leading empirical analyses to the potential presence of weak instruments; and second, it intends to help researchers to easily apply such tests in their own projects, via the `girtest` function.

Applying recently proposed tests to previous literature unveils interesting new information about the estimations and hence the conclusions drawn from them. In particular, the findings in this paper suggest that, unlike what authors originally claimed, the national legal origin indicator is not strong enough to explain the financial intermediary development, as measured by the private credit and liquid liabilities, leading to a misinterpretation of the effect of the latter on economic growth. Another remarkable issue that has arisen, in line with the recent findings in Young (2019), is the importance of the baseline assumptions to correctly interpret the estimation results: when estimating the parents' labor supply, it has shown to be crucial the parametric assumption on the structure of the error variance to be able to rely on the final estimation results. In that case, to guard oneself against misinterpretations, tests for the null of homoskedasticity against different alternatives of the error variance structure should be carried out.

On top of that, recent literature suggests that IV estimates display both high sampling uncertainty and high specification uncertainty, as minor specification changes can lead to very different estimates (see Yogo (2004), Kleibergen and Mavroeidis (2009), Mavroeidis (2010), Mavroeidis et al. (2014), Ganics (2017) or Barnichon and Mesters (2019), among others). This could turn out to be decisive in the case considered in section 4.1, where the data were averaged for the period 1960-1995, for each country. The implications would be that the strength of the legal origin indicators as instruments for the development of financial intermediaries was not constant across countries or time, but instead depended on whether we took averages for a shorter/longer period, or excluded/included some countries. More research needs to be done in this direction to disentangle such issue.

This project opens the door to future research in the field of weak instruments' tests. Further work can point, for instance, to the extension of the `girtest` package to include the cases in Ganics et al. (2018) that have not been covered here, or its application to test the statement in the previous paragraph.

References

Andrews, D., Moreira, M. and Stock, J. H. (2004), 'Optimal invariant similar tests for instrumental variables regression'.

Andrews, D. W. and Guggenberger, P. (2019), 'Identification-and singularity-robust inference for moment condition models'.

Andrews, I. (2016), 'Conditional linear combination tests for weakly identified models', *Econometrica* **84**(6), 2155–2182.

Andrews, I. and Mikusheva, A. (2016), 'Conditional inference with a functional nuisance parameter', *Econometrica* **84**(4), 1571–1612.

Angrist, J. and Evans, W. (1998), 'Children and their parents' labor supply: Evidence from exogenous variation in family size', *The American Economic Review* **88**(3), 450–477.

Barnichon, R. and Mesters, G. (2019), 'Identifying modern macro equations with old shocks', *Barcelona GSE Working Paper Series* (Working Paper n°1097).

Boyd, J. H. and Prescott, E. C. (1986), 'Financial intermediary-coalitions', *Journal of Economic theory* **38**(2), 211–232.

Breusch, T. S. and Pagan, A. R. (1979), 'A simple test for heteroscedasticity and random coefficient variation', *Econometrica: Journal of the Econometric Society* pp. 1287–1294.

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K. and Woutersen, T. (2012), 'Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments', *Econometric Theory* **28**(1), 42–86.

Cook, R. D. and Weisberg, S. (1983), 'Diagnostics for heteroscedasticity in regression', *Biometrika* **70**(1), 1–10.

Cragg, J. G. and Donald, S. G. (1993), 'Testing identifiability and specification in instrumental variable models', *Econometric Theory* **9**(2), 222–240.

Dufour, J.-M. (1990), 'Exact tests and confidence sets in linear regressions with autocorrelated errors', *Econometrica: Journal of the Econometric Society* pp. 475–494.

Dufour, J.-M. (1997), 'Some impossibility theorems in econometrics with applications to structural and dynamic models', *Econometrica: Journal of the Econometric Society* pp. 1365–1387.

Dufour, J.-M. and Jasiak, J. (2001), 'Finite sample limited information inference methods for structural equations and models with generated regressors', *International Economic Review* **42**(3), 815–844.

Ganics, G. (2017), Essays in macroeconometrics, PhD thesis, Universitat Pompeu Fabra.

Ganics, G., Inoue, A. and Rossi, B. (2018), 'Confidence intervals for bias and size distortion in IV and local projections-IV models', *Banco de España Working Paper* .

Hansen, B. E. (1999), 'The grid bootstrap and the autoregressive model', *Review of Economics and Statistics* **81**(4), 594–607.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2012), 'Instrumental variable estimation with heteroskedasticity and many instruments', *Quantitative Economics* **3**(2), 211–255.

Kent, J. T. and Hainsworth, T. J. (1995), 'Confidence intervals for the noncentral chi-squared distribution', *Journal of statistical planning and inference* **46**(2), 147–159.

Kleibergen, F. (2002), 'Pivotal statistics for testing structural parameters in instrumental variables regression', *Econometrica* **70**(5), 1781–1803.

Kleibergen, F. and Mavroeidis, S. (2009), 'Weak instrument robust tests in gmm and the new keynesian phillips curve', *Journal of Business & Economic Statistics* **27**(3), 293–311.

Levine, R., Loayza, N. and Beck, T. (2000), 'Financial intermediation and growth: Causality and causes', *Journal of monetary Economics* **46**(1), 31–77.

Mavroeidis, S. (2010), 'Monetary policy rules and macroeconomic stability: some new evidence', *American Economic Review* **100**(1), 491–503.

Mavroeidis, S., Plagborg-Møller, M. and Stock, J. H. (2014), 'Empirical evidence on inflation expectations in the new keynesian phillips curve', *Journal of Economic Literature* **52**(1), 124–88.

Montiel-Olea, J.-L. (2017), 'Admissible, similar tests: A characterization', *Unpublished manuscript* .

Montiel-Olea, J. L. and Pflueger, C. (2013), 'A robust test for weak instruments', *Journal of Business & Economic Statistics* **31**(3), 358–369.

Moreira, H. and Moreira, M. J. (2015), 'Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors', *arXiv preprint arXiv:1505.06644* .

Moreira, M. J. and Ridder, G. (2017), 'Optimal invariant tests in an instrumental variables regression with heteroskedastic and autocorrelated errors', *arXiv preprint arXiv:1705.00231* .

Porta, R. L., Lopez-de Silanes, F., Shleifer, A. and Vishny, R. W. (1998), 'Law and finance', *Journal of political economy* **106**(6), 1113–1155.

Rothenberg, T. J. (1984), 'Approximating the distributions of econometric estimators and test statistics', *Handbook of econometrics* **2**, 881–935.

Skeels, C. and Windmeijer, F. (2016), On the stock-yogo tables, Technical report, Department of Economics, University of Bristol, UK.

Staiger, D. and Stock, J. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**(3), 557–586.

Stock, J., Wright, J. H. and Yogo, M. (2002), 'A survey of weak instruments and weak identification in generalized method of moments', *Journal of Business & Economic Statistics* **20**(4), 518–529.

Stock, J. and Yogo, M. (2005), *Testing for Weak Instruments in Linear IV Regression*, Cambridge University Press, New York, pp. 80–108.

White, H. et al. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

Yogo, M. (2004), 'Estimating the elasticity of intertemporal substitution when instruments are weak', *Review of Economics and Statistics* **86**(3), 797–810.

Young, A. (2019), 'Consistency without inference: Instrumental variables in practical application', *Unpublished manuscript* .

Appendix

Here we report the results of the different approaches to test the null hypothesis of homoskedasticity against different error variance structures, in the paper by Angrist and Evans (1998) considered in section 4.2. Namely, we consider three tests that were available when the original paper was published and could have potentially been used by the authors: the ones proposed by Breusch and Pagan (1979), White et al. (1980) and Cook and Weisberg (1983).

In particular, the tests are applied to the errors of the first stage regression, which remember that read as

$$x_i = \pi_0'\mathbf{w}_i + \pi_2 s_{1i} + \gamma_0(Twoboys_i) + \gamma_1(Twogirls_i) + \eta_i \qquad \text{(1st stage)}$$

The test proposed by Breusch and Pagan (1979) tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. The first version considered assumes that the regression disturbances are independent-normal draws. The results, displayed in Table IIIa, clearly lead to reject the null hypothesis of homoskedasticity.

A more general version of the test by Breusch and Pagan (1979), suggested by Cook and Weisberg (1983), which does not restrict the disturbances to be independent-normal draws, is also used. However, even in this case we still reject the null hypothesis of homoskedasticity, as can be seen in Table IIIb.

The standard Breusch and Pagan (1979) and Cook and Weisberg (1983) tests are proven to be quite powerful in the presence of heteroskedasticity, but one may be skeptical to be restricting the analysis to too specific forms of heteroskedasticity (those of linear form). Still, if we use the White et al. (1980) test, which allows for more general (non-linear) forms of heteroskedasticity, the results are similar, as shown in Table IIIc.

Test value	2086.31
P-value	0.0000

(a) Breusch and Pagan (1979) test.

Test value	7503.19
P-value	0.0000

(b) Cook and Weisberg (1983) test.

Test value	13691.70
P-value	0.0000

(c) White et al. (1980) test.

Table III: Heteroskedasticity tests on Angrist and Evans (1998).