Gutenberg School of Management and Economics

& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

Beyond F-statistic - A General Approach for Assessing Weak Identification

Manuel Denzer (corresponding author)

Constantin Weiser

May 10, 2021

Discussion paper number 2107

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact Details:

Manuel Denzer
Johannes Gutenberg-University Mainz
Chair of Applied Statistics & Econometrics
Jakob-Welder-Weg 4
55218 Mainz, Germany
Manuel.denzer@uni-mainz.de

Constantin Weiser
Johannes Gutenberg-University Mainz
Chair of Applied Statistics & Econometrics
Jakob-Welder-Weg 4
55218 Mainz, Germany
c.weiser@uni-mainz.de

Beyond F-statistic - A General Approach for Assessing Weak Identification

Manuel Denzer *
Johannes Gutenberg-University Mainz

Constantin Weiser †
Johannes Gutenberg-University Mainz

May 10, 2021

Abstract

We propose a new method to detect weak identification in instrumental variable (IV) models. This method is based on the asymptotic normality of the distributions of the estimated endogenous variable structural equation coefficients in the presence of strong identification. Therefore, our method resulting in a specific test is more flexible than previous tests as it does not depend on a specific class of models, but is applicable for a variety of both linear and non-linear IV models or mixtures of them, which can be estimated by generalized method of moments (GMM). Moreover, our proposed test does not rely on assumptions of homoscedasticity or the absence of autocorrelation. For linear models estimated by two-stage-least-squares (2SLS), our novel test yields the same qualitative conclusions as the usually applied test on excluded instruments at the reduced form. By adopting weak identification definitions of Stock and Yogo (2005), we provide critical values for our test by means of a comprehensive Monte Carlo simulation. This enables applied econometricians to make case-by-case decisions regarding weak identification in non-homoscedastic linear models by using pair bootstrapping procedures. Moreover, we show how our insights can be applied to assess weak identification in a specific non-linear IV model.

JEL classification: C26, C36

Keywords: Weak identification, Weak instruments, Endogeneity, Bootstrap

*Gutenberg School of Management and Economics, Jakob-Welder-Weg 4, 55128 Mainz, ✉ madenzer@uni-mainz.de (corresponding author)

†Gutenberg School of Management and Economics, Jakob-Welder-Weg 4, 55128 Mainz, ✉ c.weiser@uni-mainz.de

Contents

1 Introduction

Weak instruments, or more generally weak identification, is a major concern for many IV models. In reality, detecting sufficiently strong instruments in order to explain endogenous variables can be quite difficult for many works in the field of applied econometrics. On top of being sufficiently strong, instruments have also to be exogenous. IV estimations based on weak identification can lead to biased and even inconsistent estimates (Nelson & Startz, 1990a, 1990b). This has been most prominently illustrated by Bound, Jaeger, and Baker (1995) discussing the study of Angrist and Krueger (1991), and showing that IV estimates can be inconsistent and suffer from finite-sample bias even in huge samples.

Previous research on this crucial topic for applied econometricians mainly focused on detecting weak identification in linear models. Starting with Rothenberg (1984), major contributions were made by Staiger and Stock (1997), Stock, Wright, and Yogo (2002), Hahn and Hausman (2003), Stock and Yogo (2005), Mikusheva (2013), Olea and Pflueger (2013) and I. Andrews, Stock, and Sun (2019), among others. For non-linear models, only a small scope of literature dealt with the topic of weak identification. Stock and Wright (2000), Kleibergen (2005), and I. Andrews (2018) discuss models that can be estimated by non-linear GMM. In contrast to linear models, assessing weak identification in non-linear models is more ambitious, resulting in the consequence that applied researchers tend to adopt decision rules, which are solely designed for linear models and which consequently have no theoretical foundation and validity for non-linear models. Alternatively, applied researchers tend to switch to estimate linear models, although the deficits of applying them in order to explain inherent non-linear relationships, for instance expressed by limited dependent variables, have been noticed (cf. Horrace and Oaxaca (2006)). In summary, despite all the previous progress, which we shall discuss in more detail in the next section, there is still a lack of a general decision rule that can be applied to a broad class of regression models in order to assess weak identification.

In this paper, we aim to contribute to filling this research gap by proposing a new procedure of detecting weak identification in IV models. Hence, this study complements literature on the topic of weak identification starting in the nineties. In contrast to drawing conclusions based on the examination of the conditional explanation power of the instruments on the endogenous variables by means of a linear prediction, our novel method rests on analyzing the empirical distributions of the endogenous variables' coefficients to be estimated. By gathering information on (the spread of) those distributions by means of bootstrap procedures, we propose a decision rule based on testing for deviations of the gained empirical distributions approximating their finite samples analog to the normal distribution. Therefore, our procedure which has been independently developed, shares similarities to two most recent, but unpublished, contributions (Zhan, 2017; Ievoli, 2019) and is applicable to all estimators whose distributions achieve asymptotic normality under sufficiently strong identification and asymptotic non-normality under weak identification. This is true for all GMM estimators such as classical 2SLS, limited information maximum likelihood (LIML) as well as full information maximum likelihood (FIML) estimators, for instance a (non-linear) recursive bivariate probit estimator. More precisely, we argue that

a considerable difference between the gained empirical distribution and normality is a credible indicator for weak identification, including non-identification, if minimal conditions such as a sufficient sample size hold.

As well as providing a guideline as to how our procedure can be applied to different IV models that can be estimated by GMM, we provide critical values for a corresponding test for 2SLS-estimations. This enables applied econometricians to test directly whether or not the identification of endogenous variables' coefficients in their models is sufficiently strong. The method we provide offers several advantages in comparison to those existing ones. Firstly, it is easy to apply by researchers with respect to the topic of weak identification. Secondly, it follows previous definitions in terms of considering IVs to be weak. Thirdly, it does not rest on parametric model specifications. Hence, the procedure is valid for linear as well as non-linear IV models. Fourthly, it does not require the assumption of homoscedasticity, which is a drawback of predominant test procedures but can be applied to settings characterized by heteroscedasticity. Finally, it can be extended to models characterized by multiple endogenous variables. In terms of evaluating our proposed decision rule based on the derived test, we show its analogy to classical test statistics in corresponding valid settings in terms of delivering the same qualitative conclusions. In addition, we apply it to a prominent illustration of Card (1995).

The remainder of the paper is organized as follows: Section 2 discusses different methods and tests developed in the past in order to detect weak identification in IV models. Moreover, it comprises an overview of approaches that can be seen as alternatives, given their focus on robust inference to weak identification. Section 3 presents a general IV model and contains a discussion of relevant asymptotics and test statistics for our proposed procedure. Focusing on linear models, Section 4 provides a definition of weak identification, a corresponding translation into critical values, and a formal procedure in order to assess weak identification. In addition, it comprises validations and extensions of our proposed method. In the subsequent section (Section 5), we briefly elucidate how our insights can be transferred to non-linear models and provide a specific example. Section 6 concludes and presents an outlook for further research.

2 Literature

Relevant existing literature on the topic of weak identification can be divided along two dimensions. Firstly, the proposed methods and approaches can be differentiated by being suitable solely for IV models which assume linearity in the reduced form, or conversely by being applicable also to (specific) non-linear models. While methods such as such as 2SLS - or what is known as *control function* estimator or two-stage-residual-inclusion (2SRI) (cf. Rivers and Vuong (1988); Blundell and Powell (2003)) - belong to the former sub-dimension, FIML estimators are a prominent example for the later sub-dimension. The second dimension concerns the aim of methods and approaches to detect weak identification, or alternatively to provide robust inference to weak identification. Given their limits, methods and approaches of the latter sub-dimension are of lesser interest for most applied econometric research. Among all methods

and approaches, a great majority focuses on linear models at both the reduced form as well as the structural equation. However, some of the methods and approaches contribute to more than (one part of) one dimension.

Linear Models

One of the most influential contributions in (applied) econometrics based on its citation record is the seminal study by Staiger and Stock (1997). In this study, Staiger and Stock develop an asymptotic distribution theory for a single-equation IV regression model and the 2SLS as well as the LIML estimator. Moreover, they provide measures of bias, tests of exogeneity, tests of over-identifying restrictions, as well as non-standard confidence sets for the coefficients to be estimated. Given the study's results, many researchers have concluded, and in some cases continue to conclude, incorrectly as a rule of thumb, that an F-statistic of the test on excluded instruments, also known as first stage F-test, above the value of ten ensures sufficiently strong identification in any IV setting, particularly when estimated by 2SLS.

In a sequel of this study, Stock et al. (2002) provide two different formal definitions of weak identification when discussing corresponding concerns in linear models with homoscedastic errors. Moreover, Stock et al. introduce critical values in order to enable researchers to assess weak instruments by comparing them to the above-mentioned realized F-statistic of the test on excluded instruments. However, those critical values are only valid for a selected set of linear models characterized by error distributions mentioned above.

Taking up the findings of Stock et al. (2002), Stock and Yogo (2005) discuss quantitative definitions of weak instruments in linear models more elaborately. In addition, they present critical values for testing weak identification in models estimated by a set of linear IV-estimators subject to the number of excluded instruments, the number of endogenous regressors, but also the accepted size of the bias. The set of estimators considered in this study comprises 2SLS, LIML, but also Fuller-k (Fuller, 1977) estimators. As documented by the citation record as well as the implementation into popular statistical software packages, the critical values provided by this study are the main reference for assessing weak identification in the majority of studies applying IV models besides the above-mentioned rule of thumb.[1] However, and mostly overlooked, those critical values are merely valid when specific model assumptions such as homoscedasticity are fulfilled.

A similar, relatively less influential contribution has been made by Hahn and Hausman (2003). Instead of testing on the prevalence of weak identification, Hahn and Hausman propose a procedure to test for strong instruments in a model estimated by 2SLS by means of comparing a forward and reverse regression estimator. In other words, Hahn and Hausman emphasize that the identification of the effect of the endogenous variables should be robust to a normalization of the regression in case of strong instruments. Hence, regressing the normalized dependent variable on the endogenous variables should asymptotically yield the same estimates as those when doing the opposite.

[1] The rule of thumb can actually be deduced from those critical values for the case of one endogenous variable.

More recent studies have focused on methods of detecting weak instruments in settings when assumptions raised by previous studies, such as homoskecasticity, are not fulfilled. Olea and Pflueger (2013) propose a test for detecting weak instruments which is robust to heteroscedasticity, autocorrelation, and clustering. They introduce the concept of the so-called effective F-statistic, which can be estimated as a scaled version of the non-robust F-statistic of the test on excluded instruments (Cragg & Donald, 1993), and which is identical to the robust F-statistic according to Kleibergen and Paap (2006) in case of one instrument. In addition, Olea and Pflueger provide corresponding critical values. In contrast to their contribution, our proposed method is more general since it allows more than one included endogenous regressor and can be applied to non-linear models.

While the studies mentioned above have focused mainly on the detection of weak identification in linear models, other literature has concentrated on the second dimension elucidated in the first paragraph of this section, i.e. robust inference to weak identification. In the presence of weak instruments, Mikusheva and Poi (2006) refer to three different tests which are already partly discussed in Staiger and Stock (1997) and Stock et al. (2002). More precisely, Mikusheva and Poi point out that the statistics of the proposed test by Anderson and Rubin (1949), the Lagrange multiplier (score) test (Kleibergen, 2002), and the conditional likelihood ratio test (Moreira, 2003) are robust to weak identification. Consequently, those tests can be used in order to obtain confidence regions for endogenous variables' coefficients in linear models by means of test inverting. However, those tests are limited to the cases of a single endogenous regressor and can result in unbounded or even empty confidence regions of the parameter of interest.

In a similar vein, D. Andrews and Stock (2007) as well as Mikusheva (2013) review different approaches for robust inference in linear IV models subject to weak identification. While Mikusheva focuses on extensions regarding multiple endogenous regressors, D. Andrews and Stock particularly discuss the application of the different robust tests for extensions of the standard model, such as non-normal distributed errors as well as robustness to heteroscedasticity and autocorrelation. Moreover, D. Andrews and Stock present new results for testing under many weak IV asymptotics and introduce "conditioning" methods as an alternative to the above-mentioned weak identification robust tests. Despite all the progress, each of those methods fails to fulfill the applied researchers' usual aim to provide point instead of interval estimates in IV model settings.

I. Andrews et al. (2019) provide a comprehensive overview of the different methods and approaches for a weak instrument setting in linear IV-models. They especially focus on non-homoscedastic (error) distributed data and review how studies published in the *American Economic Review* using IV regression models in the time period between 2014 and 2018 dealt with weak instruments.

Non-linear Models

Although some of the studies mentioned above, such as D. Andrews and Stock (2007), have links to or small discussions of weak identification in non-linear models, literature on those relatively more complex models is still rather incomplete, as emphasized by a small overview in

Stock et al. (2002). Previous research on the topic of weak identification in non-linear IV models focuses on models that can be estimated by non-linear GMM, where major contributions were made by the following studies.

Stock and Wright (2000) develop non-standard asymptotic approximations to the distribution of GMM estimators and test statistics in a setting of weak identification of parameters. However, those approximations are typically based on nuisance parameters, which are unknown to the researcher and which prevent direct inference. Therefore, the authors propose a procedure similar to the principle of test inverting described above. In fact, Stock and Wright suggest a method of inverting a test statistic that is directly constructed from the GMM objective function estimated by a GMM continuous updating estimator (GMM-CUE) in order to obtain confidence regions for coefficients that are robust to weak identification. However, this method cannot be considered as an unambiguous procedure to identify weak instruments in non-linear models since the confidence sets of the coefficients are jointly determined by testing weak identification as well as instrument validity.

Drawing on previous work (Kleibergen, 2002), Kleibergen (2005) proposes a GMM Lagrange multiplier (LM) statistic, which is robust to weak identification and derives its asymptotic distribution. The statistic's characteristic depends on a Jacobian estimator based on the GMM-CUE, which is asymptotically uncorrelated with the GMM moment equations. Interval estimates for the parameters of interest are established by test inverting. In contrast to the approach by Stock and Wright (2000), Kleibergen's (2005) statistic ensures that the estimated confidence regions for the parameter of interest are never empty.

I. Andrews (2018) pursues a different, but more generally applicable approach. In his study, he proposes to detect weak identification by means of constructing two-step confidence sets in GMM with controlled coverage distortions. More precisely, I. Andrews suggests to estimate identification-robust confidence sets of the parameters of interest as well as identification-non-robust confidence sets, and to check whether or not the former is contained by the latter, while allowing for some distortion. I. Andrews argues that if this condition is not fulfilled, the well-specified model suffers from weak identification with a probability tending to one.

A most recent study by Martínez-Iriarte, Sun, and Wang (2020) conflates the work by Stock and Wright (2000) and Kleibergen (2005) and provides modified corresponding test statistics. More precisely, Martínez-Iriarte et al. argue that the asymptotic distributions described by Stock and Wright and Kleibergen crucially rely on a consistent estimation of a non-parametric long-run variance estimator. Therefore, they develop fixed-smoothing asymptotics for both test statistics to account for estimation uncertainty. Those modified test statistics can be used again for the method of test inverting particularly in the presence of serial correlation.

To summarize, as documented by the respective number of studies discussed in this literature review, which only reflects the most relevant studies on the topic of weak identification, previous research has focused on linear IV models in the last twenty years. Instead of a general procedure to detect weak identification in IV models, only solutions for specific (classes of) models have

been laid out. For example, a majority of applied econometricians might be convinced in terms of sufficiently strong identification when researchers using IV-methods in linear models can present a considerably high F-statistic of the test on excluded instruments. However, a theoretical foundation which value of this F-statistic corresponds with a sufficiently strong identification in model settings which are not captured by Stock and Yogo (2005) or Olea and Pflueger (2013) does not exist. Hardly any procedures to detect weak identification exist for non-linear models and methods based on inference robust to weak identification are rarely useful, given their interval instead of point estimates.

As mentioned in the Introduction, two most recent, yet barely known contributions exist that are close to ours. Zhan (2017) constitutes a discussion paper focusing on detecting weak instruments by bootstrap procedures in linear models. By making use of the Edgeworth expansion, Zhan shows analytically that the distribution of the standardized 2SLS-estimator deviates from the standard normal distribution in case of weak instruments. For practical application, Zhan proposes to assess weak identification by using the method of residual bootstrap to derive a distribution of the standardized 2SLS-estimator and to test it against normality by making use of a simple distance test, i.e. the Kolmogorov-Smirnov test. Besides presenting a general review of the topic of weak instruments and different types of bootstrapping procedures, Ievoli (2019) presents an in-depth discussion of theoretical bootstrap asymptotics under weak instruments in his doctoral dissertation. In addition to 2SLS, he also considers other estimators, such as LIML or Fuller-k (Fuller, 1977) and other bootstrap types in comparison to the residual based one. Rather than relying on the Kolmogorov-Smirnov distance, Ievoli's proposed test of prevalent strong identification is based upon more powerful tests (Shapiro & Wilk, 1965; Jarque & Bera, 1980).

Our contribution differs to those of Zhan and Ievoli on several fronts. Firstly, both previous studies focus entirely on linear IV models and corresponding estimators. Our argumentation and derivation is based on GMM, including non-linear GMM. Although we also focus mainly on linear IV models in this study when discussing the test procedure in Section 4, the theoretical background presented in the subsequent section is relatively more general. Secondly, the specific residual based bootstrap type applied by Zhan and Ievoli prevents the consideration of heteroscedasticity and serial correlation, which is one of the key topics in linear IV models. Thirdly, the decision rule by Zhan lacks of a theoretical foundation and neither Zhan nor Ievoli provide critical values which are related to common definitions of weak identification. We shall explain the latter more detailed in Subsection 3.4.

3 Model, Asymptotics and Test Statistics

In this section, we explain the foundations for our new proposed method to assess weak identification. Firstly, we introduce a general IV model for ease of argumentation. Secondly, we discuss the asymptotic distribution of GMM estimators under strong and weak identification. Thirdly, we briefly elucidate how to obtain a sample analog of the distribution of the estimator. Fourthly,

we discuss an existing test on normality and explain how this test exploiting the gained empirical distribution of the estimator can be used for a decision on weak identification. Moreover, we show how this test statistic is related to classical and predominant statistics in order to assess weak identification in linear IV models.

3.1 General IV Model

Following the notation of Stock and Yogo (2005), this paper relies on a most general IV regression model expressed by the following system of equations:

$$\mathbf{y} = f(\mathbf{Y}\boldsymbol{\beta} + \mathbf{u}) \tag{1}$$

$$\mathbf{Y} = h(\mathbf{Z}\boldsymbol{\Pi} + \mathbf{V}) \tag{2}$$

Equation 1 constitutes the structural equation, whereas Equation 2 represents a system of reduced form equations. $\mathbf{y}$ is a $T \times 1$ vector of observations of the dependent variable where $\mathbf{Y}$ is a $T \times n$ matrix of endogenous regressors.[2] $\mathbf{Z}$ is a $T \times K$ matrix of excluded exogenous regressors, i.e. the instruments, with $\boldsymbol{\Sigma}_{\mathbf{ZZ}} = \mathbf{Z}^\intercal\mathbf{Z}/T$. $\mathbf{u}$ is a $T \times 1$ vector of structural equation errors while $\mathbf{V}$ is a $T \times n$ matrix of reduced form errors. In contrast to Stock and Yogo, we do not pose any restrictions on the errors such as homoscedasticity, but merely require that they have a mean of zero, fulfill the classical instrument related requirements of $\mathbb{E}(Zu) = \mathbb{E}(ZV) = 0$, i.e. ensuring instrument exogeneity, and are related to each other by an unspecified relationship expressed by the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_u^2 & \boldsymbol{\Sigma}_{Vu} \\ \boldsymbol{\Sigma}_{uV} & \boldsymbol{\Sigma}_{VV} \end{pmatrix} \tag{3}$$

leading to the endogeneity of $\mathbf{Y}$. Our general IV model also differentiates from that of Stock and Yogo by allowing that the dependent and endogenous variables $(\mathbf{y}, \mathbf{Y})$ both rely on a non-specific non-linear relationship, denoted by the functions f and h, on the respective regressors. $\boldsymbol{\beta}$ is the $n \times 1$ vector of parameters of interest, and $\boldsymbol{\Pi}$ constitutes a $K \times n$ matrix determining the instruments strength. In order to rule out under-identification of our model, we require $K \geq n$.

3.2 GMM Asymptotics

Newey and McFadden (1994) discuss an asymptotic distribution theory for a broad class of estimators, i.e. minimum-distance estimators. Hansen's (1982) GMM estimator comprising several other estimators, such as ordinary-least-squares (OLS), nonlinear-least-squares (NLS) or maximum likelihood (ML), is part of this class. The GMM estimator requires that a certain

[2] For ease of notation and without loss of generality, we subsume all included exogenous regressors into the vector of endogenous regressors and treat them as to be instrumented by themselves.

number of specific moment conditions are equal to zero, i.e.

$$\mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_0)] = 0, \tag{4}$$

where $g(\cdot)$ denotes a vector-valued function with its arguments $\mathbf{D}_t$ representing a multivariate random variable for each observation t out of T, i.e. the data, and the true value $\boldsymbol{\theta}_0$ of the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to estimate. Using the notation for a general IV model introduced in the previous subsection, $\mathbf{D}_t = [\mathbf{y}, \mathbf{Y}, \mathbf{Z}]$ and $\boldsymbol{\theta}_0 = \boldsymbol{\beta}_0$. By applying the empirical analog principle and the law of large numbers, the GMM estimator can be expressed as the argument minimizing a certain objective function $S_t(\boldsymbol{\theta})$, which is equal to minimizing a specific norm such as

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{arg\,min}\, S_T(\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{arg\,min} \Big(T^{-1} \sum_{t=1}^{T} g(\mathbf{D}_t, \boldsymbol{\theta}_0) \Big)^{\intercal} \hat{W}_T \Big(T^{-1} \sum_{t=1}^{T} g(\mathbf{D}_t, \boldsymbol{\theta}_0) \Big) \end{aligned} \tag{5}$$

where $\hat{W}_T$ is a consistent estimator of a positive-definite weighting matrix, i.e. $\hat{W}_T \xrightarrow{p} W$.[3]

Newey and McFadden (1994) show that this GMM estimator achieves asymptotic normality, i.e.

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N\left[0, (\mathbf{G}^{\intercal}\mathbf{W}\mathbf{G})^{-1}\mathbf{G}^{\intercal}\mathbf{W}\boldsymbol{\Omega}\mathbf{W}^{\intercal}\mathbf{G}(\mathbf{G}^{\intercal}\mathbf{W}\mathbf{G})^{-1}\right] \tag{6}$$

where $\mathbf{G} = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\, g(\mathbf{D}_t, \boldsymbol{\theta}_0)]$ and $\boldsymbol{\Omega} = \mathbb{E}[g(\mathbf{D}_t, \theta_0) g(\mathbf{D}_t, \boldsymbol{\theta}_0)^{\intercal}]$ under the following conditions: (i) $\hat{\boldsymbol{\theta}}_T$ is consistent, i.e. $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}$; (ii) $\boldsymbol{\theta}_0 \in$ interior of $\boldsymbol{\Theta}$; (iii) $g(\mathbf{D}_t, \boldsymbol{\theta}_0)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ with probability approaching to one; (iv) $\mathbb{E}[||g(\mathbf{D}_t, \boldsymbol{\theta}_0)||^2 < \infty$; (v) $\mathbb{E}[sup_{\boldsymbol{\theta} \in \mathcal{N}} ||\nabla_{\boldsymbol{\theta}}\, g(\mathbf{D}_t, \boldsymbol{\theta}_0)|| < \infty$ and (vi) the matrix $\mathbf{G}^{\intercal}\mathbf{W}\mathbf{G}$ is non-singular. As stated by Newey and McFadden, the underlying key idea is that, in large samples, estimators are approximately equal to linear combinations of sample averages. Given this circumstance, normality follows by definition of the central limit theorem (CLT).

Using the notation of Equations 1 and 2 and applying an IV estimator to a most simple model, i.e. a linear bivariate regression model ($n = 1$) with one instrument ($K = 1$) and homoscedastic errors estimated by 2SLS, the GMM moment condition is $\mathbb{E}[Z_t(y_t - Y_t\beta)] = 0$ and the asymptotic distribution of the parameter of interest $\hat{\beta}_T^{2SLS}$ can be expressed by

$$\sqrt{T}(\hat{\beta}_T^{2SLS} - \beta_0) = \frac{T^{-1/2} \sum Z_t u_t}{T^{-1} \sum Z_t Y_t} \xrightarrow{d} N\Big[0, \big(\mathbb{E}[-Z_t Y_t]\big)^{-1} \mathbb{E}[(Z_t u_t)^2] \big(\mathbb{E}[-Z_t Y_t]\big)^{-1}\Big] \tag{7}$$

where the normality follows from the normality of $T^{-1/2} \sum Z_t u_t$, which is the consequence of the CLT, as explained above. $\hat{\beta}_T^{2SLS}$ represents the classical 2SLS-estimator

$$\hat{\beta}_T^{2SLS} = \left[\mathbf{Y}^{\intercal}\mathbf{Z}(\mathbf{Z}^{\intercal}\mathbf{Z})^{-1}\mathbf{Z}^{\intercal}\mathbf{Y}\right]^{-1} \mathbf{Y}^{\intercal}\mathbf{Z}(\mathbf{Z}^{\intercal}\mathbf{Z})^{-1}\mathbf{Z}^{\intercal}\mathbf{y} = \Big[\sum_{t=1}^{T} Z_t Y_t\Big]^{-1} \sum_{t=1}^{T} Z_t y_t. \tag{8}$$

[3] For more details, see Hansen (1982).

where the last equality holds only for our specific case under investigation ($n = K = 1$).

In case of weak identification, the GMM asymptotics displayed in Equation 6 no longer apply, i.e. the difference between the parameter estimate $\hat{\boldsymbol{\theta}}$ and its true value $\boldsymbol{\theta}_0$ does not root T converge to a normal distribution. First and foremost, this is due to the inconsistency of $\hat{\boldsymbol{\theta}}$ in such a setting. This can be seen easily in the extreme case of non-identification, i.e. $\boldsymbol{\Pi} = 0$, where $\hat{\boldsymbol{\theta}}$ estimated by 2SLS in linear models converges to the probability limit of the OLS estimator, i.e. $\hat{\boldsymbol{\theta}}_T^{2SLS} \xrightarrow{p} \boldsymbol{\theta}^{OLS} = \boldsymbol{\theta}_0 + \boldsymbol{\delta} = \boldsymbol{\theta}_0 + \boldsymbol{\Sigma}_{Yu}/\boldsymbol{\Sigma}_{YY}$, where $\boldsymbol{\Sigma}_{Yu}$ denotes the covariance of $\mathbf{Y}$ and u and $\boldsymbol{\Sigma}_{YY}$ the variance of $\mathbf{Y}$. Similar behavior can be observed in non-linear models, which are estimated by NLS.[4]

Stock and Wright (2000) provide a discussion of GMM asymptotics when some or all parameters to be estimated are weakly identified.[5] Their discussion rests on some "high level" assumptions on the properties of the moments that enter the GMM first order conditions and by making use of local sequences, i.e. sequences of mean functions, to provide a non-quadratic global approximation to the objective function $S_T(\boldsymbol{\theta})$. In the most compact version: By defining $\Psi_T(\boldsymbol{\theta})$ as the centered sample moment, i.e. $\Psi_T(\boldsymbol{\theta}) = T^{-1/2}\sum_{t=1}^{T}(g(\mathbf{D}_t, \boldsymbol{\theta}_t) - \mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_t)])$, it is first assumed that $\Psi_T(\cdot) \Rightarrow \Psi(\cdot)$, where $\Rightarrow$ denotes weak convergence of random functions on $\boldsymbol{\Theta}$ with respect to the supremum norm, and where $\Psi(\cdot)$ is a Gaussian stochastic process on $\boldsymbol{\Theta}$ with mean zero and covariance function $\Omega(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}[\Psi(\boldsymbol{\theta}_1)\Psi(\boldsymbol{\theta}_2)^\intercal]$.[6] Secondly, define $m_t = T^{-1/2}\sum_{t=1}^{T}\mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_t)] = \sqrt{T}\,\mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_t)]$, which is a nonrandom (linear) mean function, uniformly converging in $\boldsymbol{\theta}$ to is limit m, i.e. $m_T(\boldsymbol{\theta}) \xrightarrow{p} m(\boldsymbol{\theta})$.[7] As before, $\hat{W}_T$ is a consistent estimator of a positive-definite weighting matrix.

Given those assumptions are fulfilled, Stock and Wright show that the GMM objective function can be rewritten as

$$
\begin{aligned}
S_T(\boldsymbol{\theta}) =& \Big(T^{-1/2}\sum_{t=1}^{T} g(\mathbf{D}_t, \boldsymbol{\theta}_0)\Big)^\intercal \hat{W}_T \Big(T^{-1/2}\sum_{t=1}^{T} g(\mathbf{D}_t, \boldsymbol{\theta}_0)\Big) \\
=& \Big(T^{-1/2}\sum_{t=1}^{T}(g(\mathbf{D}_t, \boldsymbol{\theta}_0) - \mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_0)]) + T^{-1/2}\sum_{t=1}^{T}\mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_0)]\Big)^\intercal \hat{W}_T \\
& \Big(T^{-1/2}\sum_{t=1}^{T}(g(\mathbf{D}_t, \boldsymbol{\theta}_0) - \mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_0)]) + T^{-1/2}\sum_{t=1}^{T}\mathbb{E}[g(\mathbf{D}_t, \boldsymbol{\theta}_0)]\Big) \\
=& \Big(\Psi_T(\boldsymbol{\theta}) + m_T(\boldsymbol{\theta})\Big)^\intercal \hat{W}_T \Big(\Psi_T(\boldsymbol{\theta}) + m_T(\boldsymbol{\theta})\Big)
\end{aligned}
\tag{9}
$$

[4] See Stock and Wright (2000) for more details.

[5] Please note that Stock and Wright slightly change the GMM criterion / objective function. In contrast to Hansen (1982), they multiply the GMM criterion / objective function expressed in Equation 5 by T. This does not change the optimal argument minimizing the objective function but enables some attractive asymptotic features which will be exploited. We follow Stock and Wright in the following in order to avoid any confusion.

[6] As noted by Stock and Wright, this is an extension of the functional central limit theorem (FCLT).

[7] As noted by Stock and Wright this is the GMM analog for the weak instrument asymptotics design which is reflected by $\Pi_T = C/\sqrt{T}$ in Staiger and Stock (1997) with C denoting some population constant.

which under weak identification asymptotics weakly converges to

$$S(\boldsymbol{\theta}) = \Big(\Psi(\boldsymbol{\theta}) + m(\boldsymbol{\theta})\Big)^{\intercal} W \Big(\Psi(\boldsymbol{\theta}) + m(\boldsymbol{\theta})\Big) \tag{10}$$

and consequently

$$\hat{\boldsymbol{\theta}} \Rightarrow \tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{arg\,min}\, S(\boldsymbol{\theta}). \tag{11}$$

Under weak identification (asymptotics), the part $\Psi(\boldsymbol{\theta}) + m(\boldsymbol{\theta})$ of Equation 10 follows a normal distribution, however, which does not inflate resulting in a small curvature of the objective function. This circumstance leads the objective function $S(\boldsymbol{\theta})$ following a random Chi-square type process which in turn leads to $\hat{\boldsymbol{\theta}}$ being inconsistent and having a nonstandard asymptotic distribution.[8]

For ease of understanding, consider again the above-mentioned case of the simple linear model estimated by the 2SLS-estimator of Equation 8. Assuming the most extreme version of weak identification, i.e. non-identification ($\boldsymbol{\Pi} = 0$) or in other words irrelevant instruments, the endogenous variable Y_t collapses to the error term of the reduced form V_t, leading to the following asymptotic distribution of $\hat{\beta}_T^{2SLS}$

$$\sqrt{T}(\hat{\beta}_T^{2SLS} - \beta_0) = \frac{T^{-1/2}\sum Z_t u_t}{T^{-1}\sum Z_t V_t} \xrightarrow{d} \frac{\xi_u}{\xi_v} \tag{12}$$

where due to the CLT

$$\begin{pmatrix}\xi_u \\ \xi_v\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_u^2 & \sigma_{uV} \\ \sigma_{uV} & \sigma_V^2\end{pmatrix}\right].$$

Hence, $\hat{\beta}_T^{2SLS}$ asymptotically follows a ratio of correlated normals. By means of a Cholesky factorization / linear projection, one can rewrite $\xi_u = \delta\xi_v + \xi$ where by definition $\delta = \sigma_{uV}/\sigma_V^2$, $\xi \perp \xi_v$ and ξ is normally distributed. Hence, the asymptotics of $\hat{\beta}_T^{2SLS}$ can be expressed as

$$\sqrt{T}(\hat{\beta}_T^{2SLS} - \beta_0) \xrightarrow{d} \frac{\xi_u}{\xi_v} = \delta + \frac{\xi}{\xi_v} \sim C\left[\delta, \frac{\sigma_u}{\sigma_V}\sqrt{1 - \left(\frac{\sigma_{uV}}{\sigma_u\sigma_V}\right)^2}\right] \tag{13}$$

which follows a general Gauchy distribution with a scale of $(\sigma_u/\sigma_V)\sqrt{1-(\sigma_{uV}/(\sigma_u\sigma_V)^2}$, and which is centered at the probability limit of the OLS estimator in case of irrelevant instruments $\delta = \sigma_{uV}/\sigma_V^2$ as shown above.[9] Thus, in the most extreme case of weak identification / instruments, $\hat{\beta}_T^{2SLS}$ has heavy tails and is clearly not normally distributed.

[8] See Stock and Wright (2000) for more details.
[9] See Weisstein (1999) for more details on general Gauchy distribution.

3.3 Bootstrapped Distributions

In order to derive a finite sample distribution of a parameter estimate $\hat{\boldsymbol{\theta}}$ which can be compared to its theoretical asymptotic distribution as explained in the previous subsection, we make use of resampling methods. More precisely, we apply bootstrapping methods according to Efron (1979) in order to obtain a data-based approximation for the finite sample distribution of our parameter of interest $\hat{\boldsymbol{\theta}}$.

In contrast to other contributions, we rely on the non-parametric method of pair(-wise) bootstrapping suggested by Freedman (1981) and also known as bootstrap by pairs. In general, a pair bootstrapped sample is obtained by resampling all variables, i.e. the endogenous as well exogenous variables, together from the original data. Hence, a pair bootstrapped sample consists of independent random draws of $\mathbf{D} = [\mathbf{y}, \mathbf{Y}, \mathbf{Z}]$ with replacement denoted by $\mathbf{D}^* = [\mathbf{y}^*, \mathbf{Y}^*, \mathbf{Z}^*]$. This non-parametric type of bootstrapping allows more flexibility in the comparison with other types, such as the semi-parametric residual bootstrap or even some form of parametric bootstrap. More precisely, our preferred type of bootstrapping can be applied in cases of serial correlation or heteroscedasticity, where the residual bootstrap fails to provide reliable finite sample distributions.[10] However, this comes at the cost of efficiency, i.e. our preferred bootstrap type is relatively less efficient. Nevertheless, given that we ensure a sufficiently large number of observations T in our following derivations from which our bootstrapped samples are drawn and a sufficiently large number of bootstrap replications, this deficit is negligible.

In case of strong identification in IV models that can be estimated by GMM, the bootstrap procedure is able to recover a meaningful approximation for the asymptotic distribution. Hence, given sufficiently strong identification the bootstrapped parameter estimate $\hat{\boldsymbol{\theta}}^*_T$ converges in distribution to normality, i.e.

$$\sqrt{T}(\hat{\boldsymbol{\theta}}^*_T - \boldsymbol{\theta_0}) \xrightarrow{d} N\left[0, (\mathbf{G}^\intercal \mathbf{W} \mathbf{G})^{-1} \mathbf{G}^\intercal \mathbf{W} \boldsymbol{\Omega} \mathbf{W}^\intercal \mathbf{G} (\mathbf{G}^\intercal \mathbf{W} \mathbf{G})^{-1}\right] \tag{14}$$

However, in case of weak identification, the bootstrap procedure fails to provide a good approximation. This failure of the bootstrap can be related to the failure of the Edgeworth expansion, as it has been shown in general by Hall (1992) and for the specific case of the Wald-statistic in case of weak identification by Moreira, Porter, and Suarez (2009). Exploiting the (local-to-zero) weak instrument asymptotics approach suggested by Staiger and Stock (1997), Ievoli (2019) shows how the bootstrapped distribution of the estimated parameter of interest deviates from normality for homoscedastic linear models when two different types of residual bootstrap are applied. More precisely, Ievoli shows that the bootstrapped finite sample distribution is affected by some additional components in his specific research design, which are conditional on full information on the data partly normally distributed as well as subject to randomness in case of over-identification. Consequently, test statistics, such as the Wald test for instance, cannot be asymptotically pivotal. For the specific case of non-identification ($\boldsymbol{\Pi} = 0$) as expressed by Equation 12 for the simplest

[10] In cases where serial correlation can be ruled-out, for instance by definition in cross-sectional data, the method of Wild bootstrap according to Wu (1986) can be an alternative to our preferred method.

linear model, Ievoli notes that the bootstrap completely breaks down such as asymptotic theory. This is due to fact that the true influence of the instruments, i.e. the non-identification $\mathbf{\Pi} = 0$, is never correctly estimated under this bootstrapping scheme. Therefore, the bootstrapped finite sample distribution also follows some general Gauchy distribution, but which is different from its theoretical asymptotic as expressed in Equation 12.

Zhan (2017) shows this phenomenon for linear models more generally by relying on the same asymptotic approach, i.e. the weak instrument asymptotics. Focusing on the so-called concentration parameter indicating the instruments' strength in linear models, which has to approach infinity under weak instrument asymptotics and which we shall discuss in detail in the next subsection, Zhan shows that the bootstrapping procedure does not accurately preserves the identification strength but that it preserves the pattern of identification. Moreover, Zhan shows that the bootstrap procedure does not exaggerate weak instruments problems and rather overestimates instruments strength, which can also be disadvantageous on the other side. In other words, the information conveyed by the circumstance of bootstrap failure can be exploited. This is also true when non-parametric bootstrapping techniques, such as bootstrap by pairs, are applied in contrast to the special type of residual bootstrap used by Ievoli and Zhan. As Davidson and MacKinnon (2010) suggest, the special type of residual bootstrap used in the before-mentioned studies can be seen as some variant of the bootstrap by pairs. Although this analogy cannot be seen as generalization, it can analytically be shown that the finite sample limiting distributions are the same across the two bootstrap methods.[11]

To summarize, the desirable feature of the bootstrap procedure preserving the pattern of identification yields the conclusion that a sufficiently strong deviation of the bootstrapped finite sample distribution of some IV estimator of a model that can be estimated by linear GMM in comparison to normality is a severe signal of a problem of weak identification. This is true for all estimators part of the k-class introduced by Nagar (1959), such as 2SLS, LIML or Fuller-k.

In general, there may also be reasons why bootstrap methods fail to provide a reliable finite sample approximation, which are not related to the strength of identification or the strength of instruments in linear models, and which result in invalid inference. Canty, Davison, Hinkley, and Ventura (2006) mention the presence of outliers, incorrect resampling schemes and non-pivotality as three different sources. However, we are confident that all bootstrap failure in our research setting can be attributed to problems of weak identification. In fact, given outliers, the GMM assumptions would be violated even under strong identification, which we rule out in our approach. Since we use a non-parametric resampling scheme, we are most conservative and do not apply a data generating process (DGP) leading to inhomogeneous data. In fact, the situation of non-pivotality helps us to ascertain cases of weak identification, as mentioned above.

[11] See Ievoli (2019) for corresponding expressions on the residual bootstrap (resampled instrument) type.

3.4 Test Statistics

Given the insights explained in the two previous subsections, it is natural to apply bootstrapping methods on IV models that can be estimated by GMM in order to derive an empirical finite sample distribution of the parameter of interest $\hat{\boldsymbol{\beta}}^*$, which can be tested on normality. For this reason, our proposed method of assessing weak identification is based on a test of the goodness of fit of the empirical derived distribution of the parameter of interest in comparison to its asymptotic counterpart. In contrast to relying on some general distribution tests, such as the Kolmogorov–Smirnov test, the Cramér–von Mises test or the Anderson–Darling test, we make use of the test suggested by Shapiro and Wilk (1965) and its adjusted version by Royston (1982). Given that it is explicitly designed to test for normality, it possesses the strongest power among all available tests. The test statistic under the null hypothesis, namely that the empirical distribution comes from a normally distributed population, denoted as W^{SW} in the following, is adjusted to our setting as follows

$$W^{SW}(I) = \frac{b_{\beta_T}^2}{(I-1)\,\hat{\sigma}_{\beta_T}^2} = \frac{(\sum_{i=1}^{I} a_i\,\hat{\beta}_{T(i)})^2}{\sum_{i=1}^{I}(\hat{\beta}_{Ti} - \bar{\hat{\beta}}_T)^2} \tag{15}$$

where I is the bootstrap sample size, $\hat{\beta}_{T(i)}$ is the i-th order statistic, i.e. the i-th smallest estimated coefficient in the bootstrap coefficient sample, and the vector of weights $\mathbf{a}$ can be expressed as $\mathbf{a} = (a_1, \ldots, a_I) = \left[\boldsymbol{\mu}_{(i)}^{\intercal}\boldsymbol{\Sigma}_{(i)}^{-1}\boldsymbol{\Sigma}_{(i)}^{-1}\boldsymbol{\mu}_{(i)}\right]^{-1/2}\boldsymbol{\mu}_{(i)}\boldsymbol{\Sigma}_{(i)}^{-1}$ where $\boldsymbol{\mu}_{(i)}$ is the expected value of standard normal order statistics and $\boldsymbol{\Sigma}_{(i)}$ the corresponding covariance matrix.[12]

This test statistic describes a ratio of two variances, where the denominator represents the bootstrap sample variance and the numerator contains an estimation for the bootstrap sample variance denoted as $b_{\beta_T}^2$, when the investigated distribution would follow a normal distribution. Hence, the smaller $W^{SW}(I) \in (0, 1]$, the more likely the bootstrapped empirical distribution is not normally distributed, which we consider to be a severe signal for a problem of weak identification, as discussed above.

In contrast to many other tests, the null hypothesis of the Shapiro-Wilk test on normality is rejected if the calculated test statistic is smaller than a critical value, which is tabulated for $I \leq 50$, and which can be derived by means of Monte-Carlo simulation for larger samples. Although those critical values can be seen as a naive or natural benchmark for deciding upon weak identification in IV models estimated by GMM, there is one subtle but important point. These critical values have been established and are just valid for testing on normality. They are not related to testing on weak identification. Our contribution is to link the Shapiro-Wilk test statistic to some common definition of weak identification as proposed by Stock and Yogo (2005) in order to derive new critical values for our proposed test of weak identification, which can be compared to an estimated W-statistic.[13] In other words, while an estimated W-statistic of some bootstrapped empirical distribution of $\hat{\boldsymbol{\beta}}^*$ may lead to a rejection of the null hypothesis

[12] See Shapiro and Wilk (1965) for more details.

[13] This procedure shares a similarity to Stock and Yogo (2005) which exploit a test statistic proposed by Cragg and Donald (1993) originally developed for testing under-identification.

of following a normal distribution, it might not reject the alternative null hypothesis of strong identification, because the empirical distribution is sufficiently similar to the normal distribution according to the definition of weak identification. This identification strategy is conceptually different to the previous work of Zhan (2017) and Ievoli (2019). Both do not relate their suggested statistic of testing weak identification on a common definition of weak identification, for instance in terms of the bias. More precise, while Zhan's decision rule upon weak identification is based on some arbitrary chosen cutoff value without any theoretical foundation, Ievoli sticks to the classical critical values of the Shapiro-Wilk test or those of the test proposed by Jarque and Bera (1980).

Before deriving our proposed critical values for $W^{SW}(I)$ in the next section, we discuss further advantages of our proposed method, i.e. making use of the Shapiro-Wilk test statistic on a non-parametric bootstrapped distribution of the parameter of interest $\hat{\boldsymbol{\beta}}$, in more detail. Moreover, we compare our method with test statistics, which are regularly applied in order to assess weak identification to date. As pointed out in the introduction of this paper, those statistics focus on linear models and differ with respect to their assumptions on the distribution of the errors of the model.

All of those statistics share being based on a measure proposed by Rothenberg (1984) studying weak instrument asymptotics in linear IV models with fixed instruments and Gaussian disturbances. The population concentration parameter denoted as

$$\boldsymbol{\mu}^2 = (\boldsymbol{\Sigma}_{VV}^{-1/2} \boldsymbol{\Pi}^\intercal \mathbf{Z}^\intercal \mathbf{Z} \boldsymbol{\Pi} \boldsymbol{\Sigma}_{VV}^{-1/2}) \tag{16}$$

is a unitless measure indicating the instruments' strength and collapses to

$$\mu^2 = \frac{\boldsymbol{\Pi}^\intercal \mathbf{Z}^\intercal \mathbf{Z} \boldsymbol{\Pi}}{\sigma_V^2} \tag{17}$$

in the one endogenous variable case, i.e. $n = 1$, which we will use without loss of generality for clarification purposes in the following.[14] Expressed in words, the concentration parameter measures the ratio of the variation in the endogenous variables, which can be explained by the set of instruments relative to the variation caused by the unknown error terms of the reduced form equation. The most frequently applied version of the F-statistic of the test on excluded instruments testing $\hat{\boldsymbol{\Pi}} = 0$, i.e. the relevance of the set of instruments, proposed by Cragg and Donald (1993) and taking the following expression

$$F^N = \frac{\hat{\boldsymbol{\Pi}}^\intercal \mathbf{Z}^\intercal \mathbf{Z} \hat{\boldsymbol{\Pi}} / K}{\hat{\sigma}_V^2} \tag{18}$$

is a direct estimator of the concentration parameter. This has been shown by Staiger and Stock

[14] Our motivation for reducing the dimension of $\mathbf{Y}$ to one is that the effective F-statistic proposed by Olea and Pflueger (2013), which we will introduce in the following, is merely defined for the single endogenous regressor case. The drawn insights are without loss of generality since by applying the Frisch-Waugh-Lovell Theorem (Lovell, 1963) linear models with multiple endogenous regressors can always be transferred to bivariate models. Measures for such models such as Partial F-statistics suggested by Angrist and Pischke (2009) and Sanderson and Windmeijer (2016) are based on this strategy.

(1997) by means of the relationship $\mathbb{E}(F) \cong 1 + \frac{\mu^2}{K}$ such that $F^N - 1$ can be considered as an estimator of the average instruments' strength. However, this version of the F-statistic is only valid in case of homoscedasticity. Hence, we refer to it as the non-robust version of the F-statistic.

In order to cover models characterized by absence of homoscedasticity, Kleibergen and Paap (2006) proposed a robust version of the F-statistic which is

$$F^R = \frac{\hat{\mathbf{\Pi}}^\intercal \hat{\mathbf{\Sigma}}_{\Pi\Pi}^{-1} \hat{\mathbf{\Pi}}}{K} \tag{19}$$

and which makes use of a heteroscedasticity and autocorrelation robust estimator of the standard errors of $\hat{\mathbf{\Pi}}$ according to Eicker (1967), Huber (1967), White (1980), and Liang and Zeger (1986) denoted by $\hat{\mathbf{\Sigma}}_{\Pi\Pi}$.

As already noted in Section 2, more recently, Olea and Pflueger (2013) proposed the effective F-statistic (for the single endogenous regressor case), which consists of the multiplication of the robust F-Statistic and some correction factor for non-homoscedasticity expressed as

$$F^E = \frac{\hat{\mathbf{\Pi}}^\intercal \mathbf{Z}^\intercal \mathbf{Z} \hat{\mathbf{\Pi}}}{tr(\hat{\mathbf{\Sigma}}_{\Pi\Pi}^{-1/2} \mathbf{Z}^\intercal \mathbf{Z} \hat{\mathbf{\Sigma}}_{\Pi\Pi}^{-1/2})} = \frac{K \hat{\sigma}_V^2}{tr(\hat{\mathbf{\Sigma}}_{\Pi\Pi}^{-1/2} \mathbf{Z}^\intercal \mathbf{Z} \hat{\mathbf{\Sigma}}_{\Pi\Pi}^{-1/2})} F^N \tag{20}$$

The introduction of this version of the F-statistic, which can be compared to corresponding proposed critical values, closed an important research gap for applied researchers since there is no theoretical justification to compare estimates of the robust version of the F-statistic to the critical values proposed by Stock and Yogo (2005), which is based on the non-robust version and corresponding model assumptions of homoscedasticity. As it is clearly evident, the robust and the effective F-statistic are identical in case of a single instrument, i.e. $K = 1$.

In comparison to our proposed test statistic, those different versions of the F-statistic of the test on excluded instruments exhibit several deficits. Firstly, the different versions of the F-statistic indicate the average instead of the total strength of the instrument set. Hence, in case of over-identification, i.e. $K > n$, those statistics might be misleading since they are not able to differentiate between strong and weak instruments. In other words, given all of F^R, F^N and F^E are decreasing in K, those statistics are subject to the optimality of K, which has to be decided by the researcher.[15] For example, adding irrelevant and independent instruments to a sufficiently strong instrument and assuming corresponding meaningful estimates, i.e. $\hat{\pi}_k \approx 0 \forall k > 1$, decreases the F-statistic, ultimately leading to a rejection of the test of relevant instruments despite the strong identification present in the data. Related, although delivering the same estimates for the model's coefficients and standard errors etc. at the second stage, applying the Frisch-Waugh-Lovell Theorem (Lovell, 1963) in cases of over-identification can be misleading in terms of assessing identification strength. Reducing the dimension of the set of instruments to one by means of this method leads to an automatic increase in the F-statistic by the factor K in turn, which can gauge sufficiently strong instruments even if $\mathbf{\Pi} = 0$.

[15] Please note that Equation 20 is decreasing in K because the trace in the denominator is increasing in K.

Secondly, as noted by Hahn and Hausman (2003), none of the three versions of the F-statistic considers the influence of $\mathbf{\Sigma}_{uV}$ despite its direct impact on the finite sample distribution of the 2SLS estimator among others. For the case of $n = 1$, this is shown by Hahn and Hausman (2002) expressing the bias of the 2SLS estimator approximately as a function of σ_{uV} as follows

$$\mathbb{E}[\hat{\beta}^{2SLS}] - \beta_0 \approx \frac{\sigma_{uV}}{\sigma_V^2} \frac{K}{\mu^2 + K}. \tag{21}$$

This relationship can also be seen by the following expression of the approximation of the absolute 2SLS bias proposed by Nagar (1959), which is somewhat more precise for large μ^2 in comparison to the expression by Hahn and Hausman (2002), but limited to cases of $K > 2$

$$\mathbb{E}[\hat{\beta}^{2SLS}] - \beta_0 \approx \frac{\sigma_{uV}}{\sigma_V^2} \frac{(K-2)}{\mu^2}. \tag{22}$$

Finally, all of the three different versions of the F-statistic rely on estimates of the error term (co-)variance of the reduced form equations. Hence, researchers have to make assumptions about the structure of the model's error terms, resulting in a trade-off between bias if the non-robust version of the F-statistic is applied to models with non-homoscedastic disturbances and efficiency loss if the robust or effective F-statistics are applied to models with homoscedastic disturbances.

Our proposed test statistic, i.e. the Shapiro-Wilk test W-statistic, does not suffer from most of the deficits mentioned above. The W-statistic does not dependent on the number of instruments K and any assumptions of the error term disturbances.[16] Although it is insensible to $\mathbf{\Sigma}_{uV}$, similar to the different versions of the F-statistic, we take $\mathbf{\Sigma}_{uV}$ into account following a strategy applied by Stock and Yogo (2005) when deriving critical values in the next section. In summary, there is just one small disadvantage of the W-statistic. Although less apparent, this statistic depends on the bootstrap sample size I if the bootstrapped distribution is not perfectly normally distributed, as the numerator and the denominator denoted in Equation 15 do not have to grow at the same rate when increasing I. Nevertheless, this parameter is perfectly under control for researchers in contrast to T which is exogenous. Moreover, we consider this influence in our derivation of critical values for the test of strong identification in the next section by holding I constant and providing a guideline for researchers looking for critical values when I differs from our suggestion.

4 Weak Identification Test

Taking up the idea of testing for weak identification outlined in the last section, this section presents a corresponding procedure drawing on critical values for linear models and the use of the 2SLS estimator. We restrict ourselves to this simplest of all IV estimators contained in the class of GMM estimators. However, our proposed method can also be applied to IV-estimations

[16] See Subsection 4.5.2 for some special exception in case of over-identification.

of the Fuller-k estimator.[17] Firstly, we provide a definition of weak instruments. Secondly, we derive corresponding critical values by means of a comprehensive Monte Carlo simulation. Thirdly, we propose our decision rule in order to assess weak identification. Thereafter, we relate our proposed test procedure to previously discussed test statistics of assessing weak identification by means of presenting evidence on the data-based relationship and by reporting their outcome in a prominent empirical illustration. This section concludes by providing extensions to the model used for the derivation of the critical values discussing the influence of heteroscedasticity, over-identification (for the single endogenous regressor case), and multiple endogenous explanatory variables (EEVs).

4.1 Weak Identification Sets

As first shown by Nelson and Startz (1990a, 1990b), weak identification leads to biased estimates. Therefore, we follow Stock and Yogo (2005) and define weak identification in terms of the maximum IV estimator bias. More precisely, we take over the definition by Stock and Yogo and express the bias in relative instead of absolute terms as it has already been suggested by Staiger and Stock (1997). A naive estimator not considering endogeneity, i.e. OLS for linear models, constitutes the reference. As indicated in the last section, referring to the relative bias offers the advantage that the level of endogeneity can be ignored since it is affecting both the naive estimator bias and the one considering endogeneity, but not the corresponding relative bias (Bun & Windmeijer, 2011). Stock and Yogo emphasize that the relative bias helps separating the problems of endogeneity and weak instruments. While the former is reflected by the bias in the naive estimator, the latter affects the bias of the estimator considering endogeneity.

For the leading case with $n = 1$, the relative bias reads

$$B = \frac{|\mathbb{E}[\hat{\beta}^{2SLS}] - \beta_0|}{|\mathbb{E}[\hat{\beta}^{OLS}] - \beta_0|} \tag{23}$$

where $\hat{\beta}^{2SLS}$ has to be replaced accordingly when the Fuller-k is applied instead of the 2SLS-estimator. In case of more than one endogenous variable ($n > 1$), different possibilities of how to define the relative bias exist. We comment on that in Subsection 4.5.3, but use this leading case of $n = 1$ for the following explanations and derivations without loss of generality. We follow common threshold levels for the relative bias $b \in (0.05, 0.1)$ as they have been suggested by Stock and Yogo (2005).[18] Consequently, we consider weak identification to exist if $B \geq b$ for some chosen b. This defines the weak identification set in terms of the Shapiro-Wilk test statistic as

$$\mathcal{W}_{2SLS} = \{\mathcal{Z} : B \geq b\} = \{\mathcal{Z} : W^{SW}(I) < W^{SW}_{cv}(I, b, K)\} \tag{24}$$

[17] Given that our method depends on finite sample moments, our method cannot be used for IV estimations by LIML, see Stock and Yogo (2005).

[18] In contrast to Stock and Yogo (2005), we refrain from reporting critical values for $b \in (0.2, 0, 3)$ since we do not believe that applied econometricians are willing to accept such large biases.

where $\mathcal{Z} = (\mathbf{\Pi}, \mathbf{\Sigma_{VV}}, \mathbf{\Sigma_{ZZ}}, g)$ represents a set of parameters and functional form assumptions characterizing the identification strength, and W_{cv}^{SW} denotes the critical values of the W-statistic which depend on the number of the bootstrap replications I, the tolerated relative bias b selected by the researcher, and the number of excluded exogenous regressors K, since the (asymptotic) relative bias depends on this parameter as noted by Stock and Yogo (2005).

4.2 Critical Values

The derivation of the critical values of Stock and Yogo (2005), valid for linear models with homoscedastic disturbances, crucially rests on weak instrument asymptotics. Under this process of convergence, the asymptotic expression of the relative bias denoted by Equation 23 can be linked to a weak instrument set characterized by a specific bound. This bound can then be transferred to some conservative critical value of the F-statistic of the test on excluded instruments represented by Equation 18. More precisely, Stock and Yogo show that a specific collapsing parameter Λ is effectively governing both the relative bias as well as the critical values under weak instrument asymptotics. This collapsing parameter Λ denoting the identification strength can roughly be seen as the ratio of the variance of excluded exogenous regressors multiplied with the square of the reduced form coefficients to the variance of the reduced form errors (after partialling out any covariate terms). Given specific values for n and K, Stock and Yogo use a grid-based Monte-Carlo simulation for different minimal eigenvalues of Λ to search for the minimal eigenvalue of Λ which corresponds to a relative bias b, given B is a function solely of the eigenvalues of Λ. This specific eigenvalue of Λ defining the weak instruments set is optimized by some interpolation and measures to eliminate the Monte-Carlo simulation error in order to increase precision. Using the non-central chi-squared distribution as a bounding distribution for the test statistic, which determines the Cragg and Donald (1993) statistic and follows a non-central Wishart distribution the critical values are determined by a specific percentile of the non-central chi-squared distribution divided by K, using K times the value of Λ defining the weak instruments set as parameter value for the corresponding non-centrality parameter.

Taking over this approach, which does not rest on an explicit fully specified DGP but just implicitly by the collapsing parameter Λ, in order to identify critical values is not possible in our case. This is due to the fact, that the approach by Stock and Yogo is based on the (weak instrument) asymptotic behavior of the Cragg and Donald (1993) test statistic, which is solely valid under homoscedasticity. Moreover, we are not aware of a theoretical analytic expression for the (weak instrument) asymptotic relationship between the relative bias and the W-statistic. Thus, we decide on a heuristic approach in the style proposed by Gentle (2009) in order to determine critical values which share the characteristic of those developed by Stock and Yogo, namely that the critical values of our proposed W-statistic expressed in Equation 15 are directly linked to the relative error B. Our approach is based on a DGP and a comprehensive Monte-Carlo simulation with the number of replications $R = 50000$. In the specific DGP, we hold everything constant that is known by theory not to influence the identification strength. The nuisance parameters, which have a direct influence, are sampled from meaningful distributions given their

continuous type, which prevents the use of corresponding permutations causing infinite dimension problems. The same is true for the parameters that have a direct influence on the identification strength. By making use of this approach of variability in combination with the tremendous size of the Monte-Carlo replications, we end up with a simulated dataset where we can verify that the limiting dependencies of the nuisance parameters do not affect the determination of the critical values.[19] Therefore, those can be seen as general.

Given that we focus on critical values for linear models in this section, it follows from referring to the notation of the general IV model presented in Subsection 3.1 that $f(\mathbf{Y}\boldsymbol{\beta}+\mathbf{u}) = \mathbf{Y}\boldsymbol{\beta}+\mathbf{u}$ and $h(\mathbf{Z}\boldsymbol{\Pi}+\mathbf{V}) = \mathbf{Z}\boldsymbol{\Pi}+\mathbf{V}$. As explained previously, we focus on a setting of $n = 1$ until Subsection 4.5.3 without the loss of generality. Moreover, we set $K = 1$ in the following until Subsection 4.5.2, as the (asymptotic) relative bias depends on the level of K as mentioned above. Hence, the critical values we provide in Table 1 are only valid for the simplest but also most frequent case of $n = K = 1$. We provide critical values for the most common cases of over-identification, i.e. $K > n = 1$, in Subsection 4.5.2. While included exogenous regressors have been contained in $\boldsymbol{Y}$ in our previous discussions and derivations and have been instrumented by themselves, we omit them in the following by means of the application of the Frisch-Waugh-Lovell Theorem (Lovell, 1963). Hence, all endogenous regressands and excluded exogenous regressors can be viewed to be residuals of linear projections on included exogenous regressors.

As explained above, we make use of a most flexible DGP in $R = 50000$ replications which results in any nuisance parameters having been marginalized and any limiting dependencies having been ruled out. In this DGP, we sample the number of observations from $T \sim U[100, 10000]$.[20] Based on the random sample size, we generate the excluded exogenous regressor Z, i.e. the instrument, and the structural equation error u as well as the reduced form error V as follows

$$\begin{pmatrix} Z_t \\ u_t \\ V_t \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_u^2 & \sigma_{uV} \\ 0 & \sigma_{Vu} & \sigma_V^2 \end{pmatrix}\right] \tag{25}$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_u^2 & \sigma_{uV} \\ \sigma_{uV} & \sigma_V^2 \end{pmatrix} \sim W_2(2, \mathbf{I}_2)$$

and $W_2(2, \mathbf{I}_2)$ denotes a Wishart distribution with two degrees of freedom and using an identity matrix as scale matrix. This distribution ensures that $\boldsymbol{\Sigma}$ is always positive semidefinite (PSD). In other words, while the instrument is standard normally distributed in each replication without loss of generality (and uncorrelated to the error terms fulfilling the instrument's exogeneity condition), the model errors are mean zero, yet their variance and covariance are subject to randomness. The different replications cover different degrees of endogeneity, ranging from close to uncorrelated, i.e. $\rho_{uV} = \sigma_{uV}/(\sigma_u\sigma_V) \approx 0$, up to severe correlation of close to one, i.e. $\rho_{uV} \approx 1$,

[19] Detailed analysis is available upon request.

[20] We require a minimum sample size of $T = 100$ since the bootstrap performance becomes unreliable in too small samples.

between u and V.[21]

Given those distributions, the endogenous variable Y and the outcome variable y are generated as follows

$$y_t = Y_t\beta_0 + u_t = Y_t \cdot 1 + u_t \tag{26}$$

$$Y_t = Z_t\Pi_0 + V_t \tag{27}$$

where β_0 is set to one for all replications without loss of generality. The strength of the instrument is supposed to vary, i.e. $\Pi_0 \sim U[0, 0.5]$, covering replication specific different strength of identification from irrelevant up to super strong identification, as we shall see. In fact, the identification strength per replication depends on the sampled T, Π_0 as well as σ_V^2. The expected strength of identification in our simulated dataset expressed by the concentration parameter of Equation 17 is $\mathbb{E}[\mu^2] = \mathbb{E}[\frac{\pi Z^\intercal Z \pi}{\sigma_V^2}] = \frac{0.25^2 \cdot \frac{10000+100}{2} \cdot 1}{2} = 157.81$. As mentioned-above, β is estimated by both 2SLS and OLS.

The relative bias per replication is estimated by Equation 23, where the population moments are replaced by their sample analogs. Hence, each replication setting is iterated $M = 200$ times, and the median of the $M = 200$ different values of $\hat{\beta}^{2SLS}$ ($\hat{\beta}^{OLS}$) is used as the unbiased and robust estimator for the expected value of β^{2SLS} (β^{OLS}).[22] Using the specific realized parameter set of the last of those iterations, the W-statistic denoted by Equation 15 is estimated by a bootstrap sample of size T and a number of bootstrapped replication of $I \in (99, 199, 299, 399, 499)$, leading to five different W-statistics, i.e. $W^{SW}(99), W^{SW}(199), W^{SW}(299), W^{SW}(399)$ and $W^{SW}(499)$, per single replication.[23]

Based on those $R = 50000$ simulated replications, the critical values of the W-statistics (depending on the number of bootstrap iterations I) for our test procedure presented in the subsequent subsection are extracted from the relationship between the relative bias and the estimated W-statistics. Similar to Stock and Yogo (2005), for some given tolerable relative bias b chosen by the researcher (and given a specific I), the critical values of the W-statistics are identified by means of a grid search, which ensures that the average relative bias B is not larger than the tolerated bias b. This identification can be most intuitively explained by Figure 1, which plots the data points of the estimated relative bias B and the W-statistic of $I = 499$ after excluding data points with $\rho_{uV} \leq 0.05$.[24] For some given value of tolerable bias b, for instance $b = 0.1$ as marked by the dashed red vertical line, the value of the W-statistic, which is determined by the intersection of the red dashed line with the blue line denoting the average relative bias

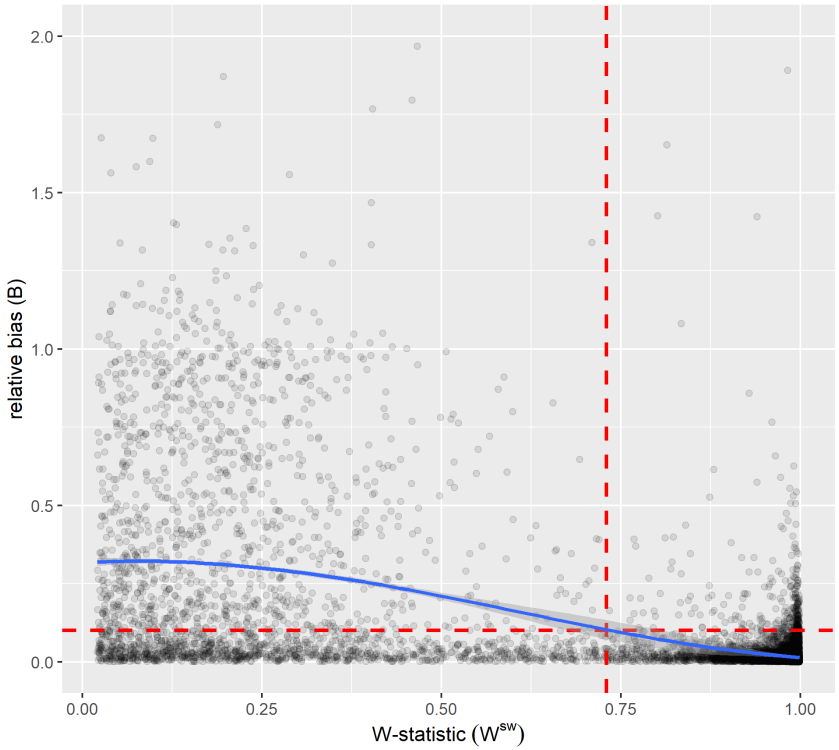
[21] See Figure A1 in the appendix for a histogram of the correlation coefficient ρ_{uV} in our simulated dataset. The relatively more frequent sampled extreme values of ρ_{uV} are caused by the Wishart distribution but not affecting our results.

[22] We use the median instead of the mean as sample analog for the expected values since it is less prone to outliers. The median is an unbiased estimator in this setting given the symmetry of the distributions of $\hat{\beta}^{2SLS}$ and $\hat{\beta}^{OLS}$.

[23] More precisely, we estimate a bootstrap sample of size 499 and randomly draw observations from this sample to calculate the W-statistics which are based on a smaller number of bootstrap replications.

[24] We exclude observations with $\rho_{uV} \leq 0.05$ since the relative bias B is an inappropriate measure for cases with small σ_{uV}. However, this restriction has only minor influence on the determination of the critical values.

Figure 1: Graphical Derivation of the Critical Values in 2SLS Estimations With $n = K = 1$



Note: For illustration purposes this figure contains only 15.000 randomly drawn data points of the total sample of $R = 50000$. However, the blue line rests on the total sample.

Table 1: Critical Values of the Test of Weak Identification in 2SLS Estimations With $n = K = 1$ and Homoscedastic Errors

	$I = 99$	$I = 199$	$I = 299$	$I = 399$	$I = 499$
$b = 0.05$	0.899	0.886	0.871	0.867	0.863
$b = 0.1$	0.838	0.795	0.766	0.749	0.737

Note: $R = 50000$, $\mathbf{\Sigma} \perp Z_t$

per different values of the W-statistic, is identified as the critical value.[25] Hence, our proposed critical value for the W-statistic of the test of sufficiently strong identification, i.e. $W_{cv}^{SW}(499)$, which is indicated by the red vertical dashed line in Figure 1, is 0.737.

Table 1 reports critical values for different values of the tolerated relative bias b and number of bootstrap iterations I (for the leading case of $n = K = 1$). As it can be seen, the critical values are intuitively decreasing in b similar to the F-statistic of Cragg and Donald (1993) used in Stock and Yogo (2005). The displayed critical values are substantially different from the classical critical values of the Shapiro-Wilk test of normality highlighting our contribution to the literature of linking the W-statistic to some measure of weak identification. The classical critical values depend on I and are above 0.974 for all different values of I displayed in Table 1.

[25] The average relative bias marked by the blue line in Figure 1 is estimated by a local regression of type locally estimated scatterplot smoothing (LOESS) with degree of smoothing equal to 0.75, degree of polynomials equal to two and least-square fitting method. See Wasserman (2006) for more details on the topic of local regression.

4.3 Decision Rule

Based on the definition of the weak identification sets in Subsection 4.1 and the derivation of corresponding critical values in the previous subsection, we suggest the following test procedure comprising the decision rule:

1. For a given dataset $\mathbf{D}_T = [\mathbf{y}, \mathbf{Y}, \mathbf{Z}]$, estimate β by means of 2SLS.
2. Choose a specific value of the number of bootstrap replications $I \in (99, 199, 299, 399, 499)$.
3. Choose a specific value of the tolerable relative bias b.
4. Apply the non-parametric pair bootstrap scheme described in Subsection 3.3 using T as bootstrap sample size yielding $\hat{\beta}_i^{*,2SLS}$.
5. Repeat Step 4. I-times and obtain a bootstrap distribution $\hat{\boldsymbol{\beta}}^{*,2SLS} = \left\{\hat{\beta}^{*,2SLS}\right\}_{i=1}^{I}$.
6. Calculate the W-statistic denoted by Equation 15 on the bootstrapped sample delivering $W^{SW,*}(I)$.
7. Decide about weak identification: If

$$W^{SW,*}(I) > W_{cv}^{SW}(I, b, K)$$

conclude that the identification is sufficiently strong, i.e. $\mathcal{Z} \notin \mathcal{W}_{2SLS}$, while if

$$W^{SW,*}(I) \leq W_{cv}^{SW}(I, b, K)$$

conclude that identification is too weak, i.e. $\mathcal{Z} \in \mathcal{W}_{2SLS}$.

If the researcher has good reasons to deviate from our proposed number of bootstrapped iterations I, we refer to our simulation code, allowing the researcher to calculate own critical values for unconsidered values of I.[26]

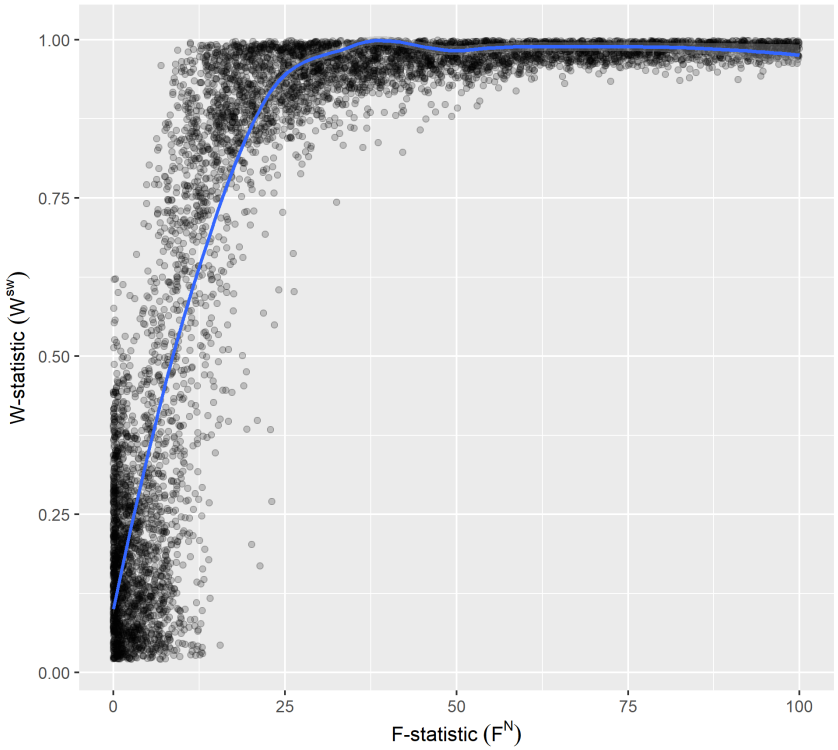
4.4 Validation

Under the DGP, and particularly the specified disturbance distributions used for the derivation of the critical values of Subsection 4.2, the non-robust F-statistic according to Cragg and Donald (1993) of the test on excluded instruments, i.e. Equation 18, is an appropriate measure for assessing weak identification. Hence, this version of the F-statistic and our proposed W-statistic should come to the same conclusions upon weak identification in such a setting.

Figure 2 visualizes the relationship between the non-robust version of the F-statistic and our proposed W-statistic using the simulated data of the derivation of the critical values. It shows

[26] The R-code is available upon request.

Figure 2: Relationship Between the Non-Robust F-Statistic and the W-Statistic



Note: For illustration purposes this figure contains only 15.000 randomly drawn data points of the total sample of $R = 50000$.

that both statistics actually deliver the same qualitative conclusions. For observations with low values of the F-statistic, the W-statistic is in the range of rejecting the null hypothesis of strong identification. On the contrary, high values of the F-statistic correspond to high values of the W-statistic. The general non-linear relationship between both statistics is marked by the blue line indicating the average value by means of a local regression.[27]

For further illustration, we apply both previously discussed test statistics and corresponding decision rules on a prominent example taken from Card (1995). Using 2SLS, Card estimates the return on education exploiting data from the National Longitudinal Survey and focusing on a cohort of young men. More specifically, Card regresses the logarithmic value of individual wage on the years of schooling and covariates where years of schooling is considered to be endogenous and instrumented by dichotomous measures of proximity to 2-year or 4-year colleges.[28] Using the original dataset of Card (1995), we follow the strategy proposed by Davidson and MacKinnon (2010) which has been taken over by Zhan (2017) and substitute the second order polynomial covariate of experience by a second order polynomial of age, since experience is considered to be endogenous, too. Moreover, we drop any of the covariates that turn out to be insignificant in the 2SLS estimation. Hence, our set of covariates consists of indicators of race, living in the south and living in standard metropolitan statistical areas besides the age polynomial, and is therefore

[27] The values of the blue line are estimated by a local regression of type LOESS with degree of smoothing equal to 0.75, degree of polynomials equal to two and least-square fitting method.

[28] The respective dummies are equal to one if there is a 2-year / 4-year college in the local labor market area the individual is living.

Table 2: Empirical Illustration: Returns to Schooling

Instrument	Proximity to 2-year college	Proximity to 4-year college
$\hat{\beta}^{2SLS}$	0.508	0.093*
SE	0.674	0.0496
F^N	0.544	10.524
W^{SW}	0.137	0.734
T	3010	3010

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Note: SE denotes the standard errors of the coefficients.

identical to Zhan (2017). Different to Card, we split up the instruments and use both indicators of proximity in separate 2SLS regressions.

Table 2 provides estimates of the returns to schooling for both instruments. It also contains the estimated non-robust F-statistic and our proposed W-statistic. According to the former, proximity to a 2-year college is a considerably weak instrument with respect to the years of schooling. However, when using proximity to a 4-year college as instrument, the test on excluded instruments yields an F-statistic, which is between the rule of thumb of Staiger and Stock (1997) and the critical value of Stock and Yogo (2005) with tolerated relative bias of $b = 0.1$. Both results perfectly match to the estimated values of the W-statistic, using the proposed test procedure of the last subsection with $I = 499$. While strong identification can clearly be rejected for the proximity to a 2-year college instrument, the estimated W-statistic of the proximity to a 4-year college specification is close to the corresponding critical value of Table 1 for $b = 0.1$.

4.5 Extensions

Given that the proposed decision rule in Subsection 4.3 relies on critical values of a most simple linear model, we discuss some extensions of the model in the following. Firstly, by altering the DGP we verify the analytically derived benefit of the W-statistic proposed in Subsection 3.4 that this test statistic does not depend on the assumption of homoscedasticity, which is a severe limitation for most methods of assessing weak identification, as discussed previously. Secondly, we analyze the impact of over-identification for the single endogenous regressor case, i.e. $K > n = 1$, and discuss the impacts of the scenario of extensive over-identification for our proposed test procedure. Thirdly, we briefly comment on how our proposed decision rule can be transferred to IV models, which can be estimated by linear GMM but which are characterized by more than one endogenous variable, i.e. $n > 1$.

4.5.1 Heteroscedasticity

As shown most recently by the replication study of Young (2020), ignoring non-homoscedastic disturbances in linear IV models can lead to severely misleading conclusions. By analyzing a

sample of 1359 instrumental variables regressions in 31 papers published in the journals of the American Economic Association, Young finds that non-homoscedastic disturbances worsen inference of all sorts. Using non-robust tests on weak instruments and the rule of thumb by Staiger and Stock (1997) leads to assess strong identification with a probability of up to 60% (depending on the concrete non-homoscedastic error process) in this sample, when the instruments are truly irrelevant. As particularized in Subsection 3.4, our proposed W-statistic and the following test procedure does not depend on correct assumptions about the error distributions. Consequently, the derived critical values in Subsection 4.2 do not depend on the assumed error specification. For instance, introducing heteroscedasticity in the DGP displayed in Equations 25, 26, 27 by means of the following distributions

$$\ddot{u}_t \sim N[0, \sigma_u^2] \tag{28}$$

$$\ddot{V}_t \sim N[0, \sigma_V^2] \cdot sin(Z_t) \tag{29}$$

and combining standardized versions of $(\ddot{u}_t, \ddot{V}_t)$ by means of a Cholesky factorization of $\mathbf{\Sigma}$ to a joint distribution of (u_t, V_t), ensuring bivariate normality under the same variance-covariance structure $\mathbf{\Sigma}$ displayed in Equation 25, but with heteroscedasticity yields critical values displayed in Table A1 for $n = K = 1$.[29] Those vary just slightly in comparison to those of Table 1 because of sampling variety.[30] However, their expected values are identical; consequently, the critical values themselves would be identical in case of $T, I \to \infty$.

Without being discussed in-depth, the logic outlined in the previous paragraph applies when the errors in a panel data IV model exhibit serial correlation instead of homoscedastic variances. Our proposed method and the corresponding test procedure are insensible to any model error disturbances.

4.5.2 Over-Identification

As emphasized in Subsection 3.4, our proposed W-statistic is independent from the number of instruments K. However, as explained in Subsection 4.1, corresponding critical values depend on K similar to Stock and Yogo (2005) as the relative bias B is a function of K. Therefore, we additionally provide critical values for settings when one endogenous variable should be explained by up to four instruments. Those settings reflecting a modest degree of over-identification are most relevant since it is already demanding to possess one valid instrument in real applied econometric settings.

In order to derive critical values of our proposed test statistic for settings with $n = 1$ and $1 < K \leq 4$, we adjust the specification of the set of instruments in our DGP displayed in Equation 25 in the sense that K independent and standard normally distributed instruments are sampled, i.e. $Z_k \sim N[0, 1] \, \forall \, k$. The reduced form equation displayed by Equation 27 is modified

[29] See the simulation R-code for additional information how the bivariate distribution of (u, V) which depends on Z but follows $N \sim [\mathbf{0}, \mathbf{\Sigma}]$ is generated. The code is available upon request.

[30] The average absolute difference amounts to 0.029.

Table 3: Critical Values of the Test of Weak Identification in 2SLS Estimations With $n = 1$ and Different K

K		$I = 99$	$I = 199$	$I = 299$	$I = 399$	$I = 499$
2	$b = 0.05$	0.968	0.969	0.967	0.966	0.966
	$b = 0.1$	0.949	0.948	0.943	0.939	0.936
3	$b = 0.05$	0.982	0.984	0.985	0.985	0.985
	$b = 0.1$	0.971	0.975	0.976	0.976	0.975
4	$b = 0.05$	0.992	0.991	0.991	0.991	0.992
	$b = 0.1$	0.979	0.984	0.986	0.986	0.986

Note: $R = 50000$ per K, $\mathbf{\Sigma} \perp Z_t$

such that $\pi_k = 0 \forall k \neq 1$. Hence, the additional $K - 1$ instruments are specified to be truly irrelevant. This parametrization and the distribution of the additional instruments are without loss of generality since, given the distribution of π_1, all different degrees of identification are still sampled in our simulated dataset.

Table 3 presents critical values for $1 < K \leq 4$, where the number of replications per different K value is the same as for Table 1, i.e. $R = 50000$. As it can be seen, the critical values are increasing in K; however, this increase is less pronounced for higher values of K. This pattern is similar to the one of the critical values proposed by Stock and Yogo (2005).

While the critical values proposed in Table 3 are meaningful for the respective parameter settings, our proposed test procedure can become invalid and misleading in specific linear over-identified models. In cases where the number of instruments is large the bootstrapped distribution of the 2SLS estimator can be asymptotically normally distributed independent of the strength of the instruments. As shown by Bekker (1994) using a sequence design where the number of instruments increases as the number of observations increases, also known as many (weak) instruments sequence / asymptotics, the 2SLS estimator becomes inconsistent but the distribution of the estimates becomes asymptotically normal although different to its limiting distribution.[31] As noted by Zhan (2017), one solution to circumvent this drawback when applying our proposed method to assess weak identification in models estimated by 2SLS is to make use of the bootstrapped distribution of a standardized version of the estimator which rest upon the estimand and standard error from conventional asymptotics. Under (improper) standardization, the bootstrapped distribution does not follow a standard normal distribution such that our proposed strategy of assessing weak identification can be applied and the bootstrapped based normality test is able to recognize weak identification.[32] Wang and Kaffo (2016) show that standard bootstrapped procedures also fail for LIML and Fuller-k estimators under many (weak) instrument sequences, but offer modified bootstrapped techniques solving this problem similar to the proposition above. In summary, given finite samples, the many (weak) instrument asymptotics can be ignored. However, in settings of models estimated by 2SLS and a

[31] The many (weak) instruments sequence can be technically expressed as in case of $n \to \infty$, $K/n \to \zeta$ where ζ describes some population constant.

[32] See Zhan (2017) for further explanation.

large set of instruments, for instance $K > 5$, we recommend that the bootstrapped distribution of the standardized estimator instead of the non-standardized estimator should be checked on normality.

4.5.3 Multiple Endogenous Variables

In case of multiple endogenous variables, i.e. $\boldsymbol{Y}$ does not collapse to a scalar, our suggested approach of detecting weak identification can be extended by means of two possibilities. As explained previously, according to the GMM asymptotics the bootstrapped distributions of the coefficients of all endogenous variables should be normally distributed in case of sufficiently strong identification. Therefore, the first option is to test normality for each bootstrapped distribution of the coefficients of the n endogenous variables separately. In this case the critical values provided in Subsection 4.2 remain valid and the single n W-statistics can be compared to them. A second possibility is to apply tests such as those proposed by Mardia (1970), Royston (1983) or Doornik and Hansen (2008), which test joint normality for the different $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_n^*)$ distributions.[33] However, firstly those tests are more restrictive in the sense that they impose and test joint normality instead of jointly marginal normal distributions. Secondly, corresponding test statistics should be compared to meaningful critical values, which depend on a joint measure of bias such as the following one proposed by Staiger and Stock (1997)

$$B = \sqrt{\frac{\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{2SLS}] - \boldsymbol{\beta}_0\right)^{\intercal}\left(\mathbf{Y}^{\intercal}\mathbf{Y}/T\right)\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{2SLS}] - \boldsymbol{\beta}_0\right)}{\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{OLS}] - \boldsymbol{\beta}_0\right)^{\intercal}\left(\mathbf{Y}^{\intercal}\mathbf{Y}/T\right)\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{OLS}] - \boldsymbol{\beta}_0\right)}} \tag{30}$$

where as before, $\mathbf{Y}$ is partialled out from any included exogenous regressors. From our perspective, there are no crucial disadvantages of the first option which, however, might lead to results of partial identification. Therefore, we refrain from providing critical values according to the relative bias displayed in Equation 30 and leave it to further research to come up with corresponding parsimonious ideas.

5 Application to Non-Linear Models

Although we have focused on linear IV models and the simplest corresponding GMM estimator, i.e. 2SLS, in the last section, our new proposed method to assess weak identification is also valid for non-linear IV models estimated by GMM, since the foundations discussed in Section 3 are not limited to linear models. In this section we briefly illustrate how our proposed method can be applied to a specific non-linear IV model estimated by FIML. More precisely, we concentrate on the case of a binary response model (BRM) with a binary EEV constituting a common setting in the fields of applied microeconomics.[34] As discussed in the literature review of Section 2, we

[33] See Henze (2002) for an overview of multivariate normality tests.

[34] See for instance leading publication such as Evans and Schwab (1995), Evans, Farrelly, and Montgomery (1999) or Altonji, Elder, and Taber (2005).

Table 4: Critical Values of the Test of Weak Identification in Recursive Bivariate Probit Estimations With $n = K = 1$

	$I = 99$	$I = 199$	$I = 299$	$I = 399$	$I = 499$
$b = 0.05$	0.988	0.991	0.993	0.994	0.994
$b = 0.1$	0.943	0.95	0.951	0.952	0.952

Note: $R = 30000$

are not aware of any existing strategy to detect weak identification in such a non-linear model setting. We apply the same procedure as in Subsection 4.2 in order to derive critical values which can then be used for our proposed test procedure on weak identification.

Taking up the general IV model presented in Subsection 3.1 and focusing on a setting without any covariate influence, a single endogenous variable and a single instrument, i.e. $K = 1$ without loss of generality, our system of non-linear equations reads

$$y_t = 1[Y_t\beta_0 + u_t > 0.5] \tag{31}$$

$$Y_t = 1[Z_t\Pi_0 + V_t > 0] \tag{32}$$

where given the binary nature of (y_t, Y_t) f and h become indicator functions represented by $1[\cdot]$. The threshold values for the latent variable expressions of Equations 31 and 32, i.e. 0 for Y_t and 0.5 for y_t, are specified to achieve an equal balance of zeros and ones for both y_t and Y_t.

Given that we still assume the errors (u_t, V_t) to be normally distributed as in Equation 25, the recursive bivariate probit estimator of Heckman (1978) and Amemiya (1978) is the natural choice of non-linear IV estimators in the class of GMM estimators due its performance benefits (Denzer, 2020). Consequently, the IV estimator is not contrasted to the OLS estimator, but to the bias of the probit estimator and the relative error becomes

$$B = \frac{|\mathbb{E}[\hat{\beta}^{Biprobit}] - \beta_0|}{|\mathbb{E}[\hat{\beta}^{Probit}] - \beta_0|}. \tag{33}$$

We follow the heuristic approach illustrated in Subsection 4.2 and take over all corresponding parametric specifications with the exception that we set σ_u^2 equal to one and that the number of replications amounts to $R = 30000$. While the former is due to the fact that $\hat{\beta}^{Biprobit}$ and $\hat{\beta}^{Probit}$ are identified up to scale and therefore have to be standardized to make them comparable across different replications, the latter is due to computation load.

Table 4 reports critical values for different values of the tolerated relative bias b and number of bootstrap iterations I for the test of weak identification in non-linear recursive bivariate probit estimations. Similar to Table 1, the critical values are decreasing in b and increasing in I.

Despite using the same approach, the critical values proposed in Table 4 are less general in comparison to those of Table 1 for two reasons. As mentioned above, coefficients are identified up to scale in non-linear models, meaning that they depend on the variance of the structural

equation's error. Since this measure of spread is unknown to the researcher in applied work, we recommend concentrating on coefficient ratios which circumvents the problem. However, this requires a modified definition of weak identification since the relative bias of the coefficient ratio is not intuitive as before. As a second point, while in linear models covariates can be partialled out by applying the theorem of Lovell (1963), this is not possible in non-linear models. We leave it to further research to take up this point and to provide adequate strategies.

6 Conclusion

In this paper, we provide a new and simple method to detect weak identification in IV models. In line with other recent but independently developed contributions (Zhan, 2017; Ievoli, 2019), our identification approach is based on making use of bootstrap procedures to derive a finite sample distribution of the structural equation coefficients of the EEVs in IV models. Given some minimal requirements such as a sufficient sample size to obtain meaningful bootstrap distributions hold, deviations of those empirical finite sample distributions to normality can be considered as severe signal for weak identification. In contrast to previous test statistics to detect weak identification, which are almost exclusively limited to linear models and assumptions of homoscedasticity, our proposed method is applicable to all IV models that can be estimated by GMM. Hence, our proposed method covers models with non-homoscedastic disturbances, but also models estimated by non-linear GMM such as FIML.

As well as discussing the theoretical background of our proposed method, i.e. GMM asymptotics (under weak identification), we contribute to literature by providing critical values based on an exhaustive Monte-Carlo simulation for a complete test procedure based on our proposed method. This enables applied econometricians to test for weak identification in their research settings of linear models with endogenous regressors estimated by the most common estimator, i.e. 2SLS, when classical tests of weak identification cannot be used or when researchers do not want to make use of weak identification robust inference methods leading to interval instead of point estimates for parameters of interest. Our proposed test procedure is easy to apply and is linked to the definition of weak identification used in the seminal paper of Stock and Yogo (2005), as it exploits the same and intuitive metric relating the bias of the IV estimator to the bias of some naive estimator ignoring endogeneity. We validate our test procedure by investigating its performance in a prominent empirical illustration. Moreover, we show that our test procedure delivers the same qualitative results when applied to settings where classical tests of weak identification are valid. While we focus on the leading linear IV model case with a single endogenous regressor and a single instrument, we discuss how our proposed method can be applied to settings with more than one instrument for a single endogenous variable, which constitutes the most frequent setting of over-identification, and to settings of multiple endogenous variables. For a specific non-linear model, we briefly illustrate how our proposed test procedure can be applied.

Further research for refining our proposed method could concentrate on several issues. Firstly, it would be beneficial if the test statistic of the used normality test in our proposed procedure

could be analytically linked to the relative bias, as is the case for the F-statistic of Cragg and Donald (1993). This would make our applied heuristic approach obsolete and provide additional precision. Although the Shapiro-Wilk test statistic offers advantages such as its power and its boundedness, it is not trivial to find an analytical relationship to the relative bias. Secondly and possibly based on improvements with respect to the former, future research could verify our test procedure by means of assessing its power. Corresponding supporting outcome would further confirm the validity of our suggested approach. Thirdly, it would be beneficial to overcome the two loose ends with respect to the application of our proposed method to non-linear IV models estimated by GMM, namely the limits with respect to the relative bias definition and the impact of multiple explanatory variables, as outlined in the previous section. Finally, it could be interesting to apply our proposed procedure to a weakly endogenous instrument setting according to Ievoli (2019) in order to investigate the outcome when the instruments' requirement on exogeneity instead of relevance is not fulfilled.

Despite this scope of improvement, this paper supports a new stream in econometrics to think about the detection of weak identification particularly in non-classical models, and to provide reliable and more general metrics. Being the first in this regard, it provides a concrete test procedure which is linked to well-established criteria in the field of econometrics, and which should help applied researchers in their work.

References

Altonji, J., Elder, T., & Taber, C. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, *113*(1), 151-184. doi: 10.1086/426036

Amemiya, T. (1978). The Estimation of a Simultaneous Equation Generalized Probit Model. *Econometrica*, *46*(5), 1193–1205.

Anderson, T., & Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63.

Andrews, D., & Stock, J. (2007). Inference with Weak Instruments. In R. Blundell, W. Newey, & T. Persson (Eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol III.* Cambridge: Cambridge University Press. (Prepared for the 2005 World Congress of the Econometric Society)

Andrews, I. (2018). Valid Two-Step Identification-Robust Confidence Sets for GMM. *The Review of Economics and Statistics*, *100*(2), 337-348.

Andrews, I., Stock, J., & Sun, L. (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, *11*(1), 727-753.

Angrist, J., & Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Angrist, J., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Bekker, P. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, *62*(3), 657–681.

Blundell, R., & Powell, J. (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. In M. Dewatripont, L. Hansen, & S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress* (Vol. 2, p. 312-357). Cambridge University Press. doi: 10.1017/CBO9780511610257.011

Bound, J., Jaeger, D., & Baker, R. (1995). Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Bun, M., & Windmeijer, F. (2011). A Comparison of Bias Approximations for the Two-Stage Least Squares (2SLS) Estimator. *Economics Letters*, *113*(1), 76-79. doi: 10.1016/j.econlet.2011.05.047

Canty, A., Davison, A., Hinkley, D., & Ventura, V. (2006). Bootstrap Diagnostics and Remedies. *Canadian Journal of Statistics*, *34*(1), 5-27. doi: 10.1002/cjs.5550340103

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In L. Christofides, E. Grant, & R. Swidinsky (Eds.), *Aspects of Labor Market Behavior: Essays in Honour of John Vandercamp.* Toronto: University of Toronto Press.

Cragg, J., & Donald, S. (1993). Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, *9*(2), 222–240.

Davidson, R., & MacKinnon, J. (2010). Wild Bootstrap Tests for IV Regression. *Journal of Business & Economic Statistics*, *28*(1), 128-144. doi: 10.1198/jbes.2009.07221

Denzer, M. (2020). *Estimating Causal Effects in Binary Response Models with Binary Endogenous Explanatory Variables - A Comparison of Possible Estimators.*

Doornik, J., & Hansen, H. (2008). An Omnibus Test for Univariate and Multivariate Normality. *Oxford Bulletin of Economics and Statistics*, *70*(s1), 927-939. doi: 10.1111/j.1468-0084.2008.00537.x

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Eicker, F. (1967). *Limit Theorems for Regressions with Unequal and Dependent Errors.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp. 59–82.

Evans, W., Farrelly, M., & Montgomery, E. (1999). Do Workplace Smoking Bans Reduce Smoking? *American Economic Review*, *89*(4), 728-747. doi: 10.1257/aer.89.4.728

Evans, W., & Schwab, R. (1995). Finishing High School and Starting College: Do Catholic Schools Make a Difference? *The Quarterly Journal of Economics*, *110*(4), 941-974. doi: 10.2307/2946645

Freedman, D. (1981). Bootstrapping Regression Models. *The Annals of Statistics*, *9*(6), 1218 – 1228. doi: 10.1214/aos/1176345638

Fuller, W. (1977). Some Properties of a Modification of the Limited Information Estimator. *Econometrica*, *45*(4), 939–953.

Gentle, J. (2009). *Computational Statistics.* Dordrecht: Springer. doi: 10.1007/978-0-387-98144-4

Hahn, J., & Hausman, J. (2002). Notes on Bias in Estimators for Simultaneous Equation Models. *Economics Letters*, *75*(2), 237-241. doi: 10.1016/S0165-1765(01)00602-4

Hahn, J., & Hausman, J. (2003). Weak Instruments: Diagnosis and Cures in Empirical Econometrics. *American Economic Review*, *93*(2), 118-125.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag New York.

Hansen, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, *50*(4), 1029–1054.

Heckman, J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, *46*(4), 931–959.

Henze, N. (2002). Invariant Tests for Multivariate Normality: A Critical Review. *Statistical Papers*, *43*(4), 467-506. doi: 10.1007/s00362-002-0119-6

Horrace, W., & Oaxaca, R. (2006). Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model. *Economics Letters*, *90*(3), 321 - 327. doi: 10.1016/j.econlet.2005.08.024

Huber, P. (1967). *The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp. 221–233.

Ievoli, R. (2019). *Essays on Bootstrap Inference Under Weakly Identified Models* (Unpublished doctoral dissertation). Università di Bologna.

Jarque, C., & Bera, A. (1980). Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters*, *6*(3), 255-259. doi: 10.1016/0165-1765(80)90024-5

Kleibergen, F. (2002). Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression. *Econometrica*, *70*(5), 1781–1803.

Kleibergen, F. (2005). Testing Parameters in GMM Without Assuming That They Are Identified. *Econometrica*, *73*(4), 1103-1123.

Kleibergen, F., & Paap, R. (2006). Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics*, *133*(1), 97-126. doi: 10.1016/j.jeconom.2005.02.011

Liang, K.-Y., & Zeger, S. (1986, 04). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, *73*(1), 13-22. doi: 10.1093/biomet/73.1.13

Lovell, M. (1963). Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. *Journal of the American Statistical Association*, *58*(304), 993-1010. doi: 10.1080/01621459.1963.10480682

Mardia, K. (1970, 12). Measures of Multivariate Skewness and Kurtosis With Applications. *Biometrika*, *57*(3), 519-530. doi: 10.1093/biomet/57.3.519

Martínez-Iriarte, J., Sun, Y., & Wang, X. (2020). Asymptotic F-Tests Under Possibly Weak

Identification. *Journal of Econometrics*, *218*(1), 140 - 177. doi: 10.1016/j.jeconom.2019.10.011

Mikusheva, A. (2013). Survey on Statistical Inferences in Weakly-Identified Instrumental Variable Models. *Applied Econometrics*, *29*(1), 117-131.

Mikusheva, A., & Poi, B. (2006). Tests and Confidence Sets With Correct Size When Instruments are Potentially Weak. *The Stata Journal*, *6*(3), 335-347.

Moreira, M. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, *71*(4), 1027-1048. doi: 10.1111/1468-0262.00438

Moreira, M., Porter, J., & Suarez, G. (2009). Bootstrap Validity for the Score Test When Instruments may be Weak. *Journal of Econometrics*, *149*(1), 52-64. doi: 10.1016/j.jeconom.2008.10.008

Nagar, A. (1959). The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica*, *27*(4), 575–595.

Nelson, C., & Startz, R. (1990a). The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument is a Poor One. *The Journal of Business*, *63*(1), 125–140.

Nelson, C., & Startz, R. (1990b). Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator. *Econometrica*, *58*(4), 967–976.

Newey, W., & McFadden, D. (1994). Chapter 36 Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics* (Vol. 4, p. 2111 - 2245). Elsevier. doi: 10.1016/S1573-4412(05)80005-4

Olea, J., & Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, *31*(3), 358-369. doi: 10.1080/00401706.2013.806694

Rivers, D., & Vuong, Q. (1988). Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models. *Journal of Econometrics*, *39*(3), 347 - 366. doi: 10.1016/0304-4076(88)90063-2

Rothenberg, T. (1984). Chapter 15 Approximating the Distributions of Econometric Estimators and Test Statistics. In *Handbook of Econometrics* (Vol. 2, p. 881 - 935). Elsevier. doi: 10.1016/S1573-4412(84)02007-9

Royston, J. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(2), 115–124.

Royston, J. (1983). Some Techniques for Assessing Multivarate Normality Based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *32*(2), 121–133.

Sanderson, E., & Windmeijer, F. (2016). A Weak Instrument F-Test in Linear IV Models With Multiple Endogenous Variables. *Journal of Econometrics*, *190*(2), 212-221. doi: 10.1016/j.jeconom.2015.06.004

Shapiro, S., & Wilk, M. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611.

Staiger, D., & Stock, J. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, *65*(3), 557–586.

Stock, J., & Wright, J. (2000). GMM With Weak Identification. *Econometrica*, *68*(5), 1055-1096.

Stock, J., Wright, J., & 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., & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In D. Andrews (Ed.), *Identification and Inference for Econometric Models* (p. 80-108). New York: Cambridge University Press.

Wang, W., & Kaffo, M. (2016). Bootstrap Inference for Instrumental Variable Models With Many Weak Instruments. *Journal of Econometrics*, *192*(1), 231-268. doi: 10.1016/j.jeconom.2015.12.016

Wasserman, L. (2006). *All of Nonparametric Statistics*. Dordrecht: Springer. doi: 10.1007/0-387-30623-4

Weisstein, E. (1999). *CRC Concise Encyclopedia of Mathematics* (1st ed.). Chapman & Hall: CRC Press.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Wu, C. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, *14*(4), 1261 – 1295. doi: 10.1214/aos/1176350142

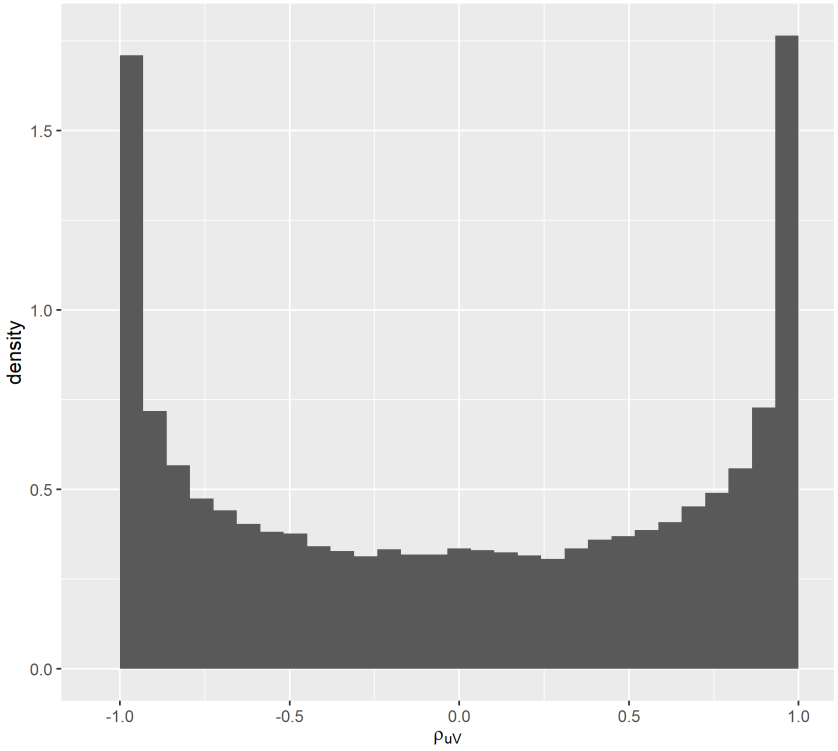
Young, A. (2020). *Consistency without Inference: Instrumental Variables in Practical Application.*

Zhan, Z. (2017). *Detecting Weak Identification by Bootstrap.*

Appendix

Figures

Figure A1: Histogram of the Correlation Coefficient ρ_{uV} in the Simulation Sample



Note: $R = 150000$

Tables

Table A1: Critical Values of the Test of Weak Identification in 2SLS Estimations With $n = K = 1$ and Heteroscedastic Errors

	$I = 99$	$I = 199$	$I = 299$	$I = 399$	$I = 499$
$b = 0.05$	0.888	0.874	0.844	0.846	0.825
$b = 0.1$	0.813	0.771	0.725	0.710	0.682

Note: $R = 50000$, $\mathbf{\Sigma} \not\perp Z_t$