Dynamic Identification Using System Projections on Instrumental Variables*

Daniel J. Lewis
University College London

Karel Mertens†
FRB Dallas, CEPR

July 1, 2024

Abstract

We propose System Projections on Instrumental Variables (SP-IV) to estimate structural relationships using regressions of structural impulse responses obtained from local projections or vector autoregressions. Relative to IV with distributed lags of shocks as instruments, SP-IV imposes weaker exogeneity requirements and can improve efficiency and increase effective instrument strength relative to the typical 2SLS estimator. We describe inference under strong and weak identification. The SP-IV estimator outperforms other estimators of Phillips Curve parameters in simulations. We estimate the Phillips Curve implied by the main business cycle shock of Angeletos et al. (2020) and find that the impulse responses are consistent with weak but also relatively strong cyclical connections between inflation and unemployment.

JEL classification: E3, C32, C36.

Keywords: Structural Equations, Instrumental Variables, Impulse Responses, Robust Inference, Phillips Curve, Inflation Dynamics.

*The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System. We are grateful to Kirill Borusyak, Xu Cheng, Andrew Chesher, Richard Crump, Marco Del Negro, Oscar Jordà, Dennis Kristensen, Geert Mesters, José Luis Montiel Olea, Mikkel Plagborg-Møller, Andrei Zeleneev, and seminar and conference participants at UPenn, Penn State, University of Houston, UCL, UPF, Oxford, Johns Hopkins, Boston University, Padova, Bologna, NBER/NSF Time Series, LSE, Manchester, University of Namur, Singapore University, the National Bank of Belgium, the Bank of Canada, KU Leuven Macro Workshop, NBER Summer Institute, EABCN Conference on Local Projections and Central Banking, Bank of England, and the Federal Reserve Banks of Kansas City and New York.

†*Contact*: mertens.karel@gmail.com, tel: +(214) 922-6000

This paper studies the estimation of β in structural time series equations of the form

$$y_t = \beta' Y_t + u_t \ , \tag{1}$$

where y_t is a scalar observation of an outcome variable in period t, Y_t is a $K \times 1$ vector of explanatory variables, u_t is an error term (which may or may not be i.i.d.), and β contains the K structural parameters of interest. The explanatory variables Y_t may contain contemporaneous variables but also lagged variables or agents' expectations of future variables that the econometrician may not measure well. We are interested in applications where $E[Y_t u_t] \neq 0$, such that standard regression techniques yield inconsistent estimates of β due to endogeneity.

Equation (1) nests a wide range of dynamic relationships of interest in macroeconomics. Consider the example of the Hybrid New Keynesian Phillips Curve (henceforth, the "Phillips Curve"),

$$\pi_t = \gamma_b \pi_{t-1} + \gamma_f \pi^e_{t+1} + \lambda gap_t + u_t \ , \tag{2}$$

where π_t denotes inflation, π^e_{t+1} is a measure of price setters' period t expectation of inflation in $t+1$, and gap_t is an output gap measure (the deviation of actual economic activity from the level without price rigidities). Equation (2) maps into the more general problem in (1) with $y_t = \pi_t$, $Y_t = [\pi_{t-1}, \pi^e_{t+1}, gap_t]'$ and $\beta = [\gamma_b, \gamma_f, \lambda]'$. The estimation of β is complicated by a number of well-known problems that result in $E[Y_t u_t] \neq 0$, see for instance Mavroeidis et al. (2014), McLeay and Tenreyro (2019) or Barnichon and Mesters (2020). One source of endogeneity is *measurement error*, as in practice the output gap and inflation expectations must be replaced with proxy measures. A second source of endogeneity is *simultaneity* since the error term generally includes structural shocks that also influence the endogenous variables in Y_t. Many theoretical dynamic relationships include expectations and other endogenous explanatory variables and, therefore, face similar problems.

A common approach in the literature is to rely on dynamics for identification and use lagged variables as instrumental variables. In the Phillips

Curve application, it is, for instance, typical to use $gap_{t-1}, gap_{t-2}, \ldots$ and $\pi_{t-2}, \pi_{t-3}, \ldots$, or lags of other readily available macroeconomic variables.[1] Instrument exogeneity, in this case, requires that the error term u_t is uncorrelated with any of the instrumenting lagged macroeconomic variables. In other words, the shocks (and lags thereof) influencing the error term u_t must be uncorrelated with the shocks influencing the lagged variables used as instruments. There is no general reason to believe that restrictions of this sort hold when the instruments are lags of standard macroeconomic variables since the latter may be functions of the same past shocks that determine the error term. Lags of the output gap or inflation are, for example, not valid instruments for (2) in medium-scale macroeconomic models such as the Smets and Wouters (2007) model.

For this reason, Barnichon and Mesters (2020) propose IV with current and lagged values of external measures of monetary policy shocks as instruments, as these measures are potentially more credibly uncorrelated with u_t than lagged macro variables. In general, however, the literature is rarely comfortable with imposing the strong assumption of unconditional lag exogeneity on available external measures of structural shocks and typically avoids doing so by including a rich set of lagged macroeconomic controls in vector autoregressive models (VARs) and local projections (LPs). Unfortunately, when estimating structural equations rather than impulse responses, including such controls in conventional single-equation IV (SE-IV) regressions with a distributed lag (DL) of shocks as instruments shrinks the explanatory power of the instruments to that of only the contemporaneous shock, resulting in weaker or even under-identification. A researcher may, for example, wish to follow the recommendations of Stock and Watson (2018) and include lags of the external shock measure as controls when estimating structural impulse response functions (IRFs) using local projections. However, including those lags as controls is not possible when they are used as instruments as in Barnichon and Mesters (2020). Similar problems arise when the controls do not in-

[1] For example, in an influential paper, Galí and Gertler (1999) use four lags of inflation, the labor income share, the output gap, the long-short interest rate spread, wage inflation, and commodity price inflation. They treat π_{t-1} as exogenous, and so also use it as an instrument.

clude lags of the shock measure but contain variables that are correlated with the lagged shock measures.

In this paper, we propose a novel approach to identifying and estimating β that allows the inclusion of lagged variables as controls without weakening identification. Specifically, we replace the single equation (1) with an H-dimensional system of structural equations in forecast errors of y_t and Y_t, where H is the number of leads. The forecast errors can be derived from a variety of forecasting models, including VARs or LPs with a rich set of controls. The contemporaneous values of the N_z instrumental variables generate HN_z moment conditions, which we solve in closed form for β, yielding a restricted IV estimator in the system of reduced form forecast errors. We refer to this methodology as System Projections on Instrumental Variables, or SP-IV.

SP-IV estimates structural equations on the basis of the relationships between empirical estimates of the dependent and independent variables' impulse responses to economic shocks. We show that SP-IV is equivalent to a straightforward regression of the IRF of y_t on the IRFs of Y_t, where the IRFs can be obtained from a VAR, LPs, or other valid impulse response estimators. Intuitively, SP-IV finds the linear combination of IRFs of the endogenous variables to one or more suitably chosen structural shocks that most closely matches the IRFs of the dependent variable to the same shocks. Moreover, these IRFs can be obtained from any LP or VAR identification scheme and – unlike the SE-IV approach of Barnichon and Mesters (2020) – SP-IV, therefore, does not require the availability of external shock measures.

Depending on the data generating process (DGP), SP-IV also has several further advantages relative to SE-IV using a DL of instruments. First, it can leverage existing external shock measures just like SE-IV, but with adequate controls, it requires only the weaker assumptions of contemporaneous and lead exogeneity of the instruments, compared to contemporaneous, lead, and lag exogeneity for SE-IV. Second, the use of forecast errors instead of raw variables can improve efficiency in estimating β relative to the typical 2SLS implementation of SE-IV. Third, similar efficiency gains in the first stage can increase effective instrument strength, thereby mit-

igating weak instrument problems. The latter two advantages of SP-IV are more likely when the error term/endogenous variables are more persistent/predictable and are, therefore, likely to be relevant in many macro time series applications.

As SP-IV is a GMM estimator, inference is straightforward under strong identification. We develop a first-stage test for instrument strength by extending the popular bias-based test in Stock and Yogo (2005) to the SP-IV setting. As instruments are often weak in practice, we propose weak instrument robust inference procedures based on the Anderson and Rubin (1949) AR statistic and Kleibergen's (2005) KLM statistic.

We demonstrate the potential performance gains of SP-IV in simulations estimating the Phillips curve parameters using data generated from the Smets and Wouters (2007) model. When the instrument is lag endogenous, 2SLS with DL instruments is prohibitively biased, but SP-IV with suitable (but realistic) controls is not. When the instruments are valid for both estimators, SP-IV with controls exhibits considerably smaller bias than 2SLS with DL instruments in finite samples. A VAR implementation of SP-IV has the lowest bias of all estimators we consider, while LP implementations have lower variance. Robust inference procedures for SP-IV control size in realistic sample sizes and exhibit smaller size distortions when HN_z is large than similar procedures for conventional SE-IV.

As an empirical application, we estimate the Phillips curve in US data using the Main Business Cycle (MBC) shock of Angeletos et al. (2020) as an instrument. Identified as the shock that maximally explains the cyclical variation in unemployment, Angeletos et al. (2020) conclude from its muted impact on inflation that the Phillips curve must be very flat. However, we find that robust confidence sets for the slope of the Phillips curve are consistent with weak but also fairly strong cyclical connections between inflation and economic activity. This application illustrates how SP-IV enables formal assessments of structural relationships between IRFs over multiple horizons while accounting for the sampling error in the IRF estimates. After properly accounting for estimation uncertainty, the evidence from IRFs to an MBC shock does not provide strong support for inflation dynamics that are disconnected from the business cycle.

Researchers frequently draw conclusions about structural economic relationships by looking at relative magnitudes of IRF coefficients. However, only a few existing studies use regressions with IRFs to more formally estimate the parameters in these relationships. An early contribution by Jordà and Kozicki (2011) proposes a regression of IRFs from LPs as a solution to a minimum distance problem to identify structural parameters. However, Jordà and Kozicki (2011) abstract from issues of identification and consider only reduced-form IRFs, relying on a high-level assumption for consistency. Our paper instead develops theory for regressions with "structural" IRFs from both LPs or VARs. More recently, Barnichon and Mesters (2020) show that 2SLS with a DL of an economic shock as instruments is equivalent to a regression with IRFs estimated from DL regressions. The SP-IV approach in this paper allows the IRFs to come from general VAR or LP specifications and identification schemes, and we demonstrate several other advantages of SP-IV relative to 2SLS with DL instruments. Inspired by Barnichon and Mesters (2020), Del Negro et al. (2020) regress posterior draws of impulse response coefficients from a Bayesian VAR to estimate Phillips Curve parameters, but they do not provide a theoretical development of their method; in particular, they do not relate it to IV. Finally, Galí and Gambetti (2020) identify markup shocks in a VAR and estimate the Phillips curve parameters using counterfactual data generated from the VAR after setting all realizations of the markup shocks to zero. This approach can also be viewed as a regression with IRFs, in this case to all relevant shocks except the markup shock that causes simultaneity problems.

Henceforth, $\otimes$ denotes the Kronecker product, $\text{Tr}(\cdot)$ the trace operator, $\text{vec}(\cdot)$ the vectorization operator, $\text{mineval}\{\cdot\}$ /$\text{maxeval}\{\cdot\}$ the minimum/maximum eigenvalue, $\xrightarrow{p}$ convergence in probability, $\xrightarrow{d}$ convergence in distribution, and $P_X = X'(XX')^{-1}X$ the projection matrix.

1. System Projections on Instrumental Variables

We begin by reformulating the dynamic relationship in (1) in terms of forecast errors. Taking h-horizon leads and taking residuals after condi-

tioning on an $N_x \times 1$ vector of predetermined variables X_{t-1} yields

$$y_t^{\perp}(h) = \beta' Y_t^{\perp}(h) + u_t^{\perp}(h) \ , \tag{3}$$

where $y_t^{\perp}(h) = y_{t+h} - E[y_{t+h} \mid X_{t-1}]$, $Y_t^{\perp}(h) = Y_{t+h} - E[Y_{t+h} \mid X_{t-1}]$, and $u_t^{\perp}(h) = u_{t+h} - E[u_{t+h} \mid X_{t-1}]$. Let z_t denote an $N_z \times 1$ vector of instrumental variables, and define $z_t^{\perp} = z_t - E[z_t \mid X_{t-1}]$. As explained in the introduction, we focus on applications that rely on dynamics for identification, exploiting orthogonality conditions between the error term u_t and $z_t, z_{t-1}, ...$ Instead of the usual approach of imposing orthogonality between z_{t-h} and u_t for various $h \geq 0$, we impose

$$E[u_t^{\perp}(h) z_t^{\perp}] = 0 \ ; \ h = 0, \ldots, H-1 \ . \tag{4}$$

Without conditioning on X_{t-1} and under stationarity, the orthogonality conditions in (4) are equivalent to imposing $E[u_t z_{t-h}] = 0$, as in conventional SE-IV. The key departure compared to using a DL of z_t as instruments is that the moments in (4) are not in terms of the unconditional data but in terms of forecast errors after conditioning on the predetermined predictors X_{t-1}, where lags of z_t may be included in X_{t-1}.

1.1. The Generalized Method of Moments Problem

The conditions in (4) provide a set of HN_z moment conditions that can be used to identify the K elements of β. Let $y_{H,t}^{\perp}$ and $u_{H,t}^{\perp}$ denote the $H \times 1$ vectors in which the $(h+1)$-th element is $y_t^{\perp}(h)$ or $u_t^{\perp}(h)$ respectively. Let $Y_{H,t}^{\perp}$ denote the $HK \times 1$ vector stacking the $H \times 1$ vectors $Y_{H,t}^{k,\perp}$, where Y_t^k is the k-th variable in Y_t. Using this notation, the moment conditions are

$$E[u_{H,t}^{\perp}(\beta) \otimes z_t^{\perp}] = 0 \ , \tag{5}$$

where $u_{H,t}^{\perp}(b) \equiv y_{H,t}^{\perp} - (b' \otimes I_H) Y_{H,t}^{\perp}$ and the truth is $b = \beta$. Note that, after expressing $u_t^{\perp}(h)$ explicitly as a function of β, this expression simply stacks (4) across horizons.

The moment conditions in (5) can be augmented to account for the estimation of the forecast errors. We consider the class of forecasting

models for the conditional expectations underlying (3) that are linear in X_{t-1} but possibly nonlinear in a set of parameters collected in the vector d. This class includes LPs and VARs, both of which are widely used in applied macroeconomics.[2] The forecasting moment conditions are

$$E\left[X_{t-1} \otimes \left[y_{H,t}^{\perp\prime}(\zeta), Y_{H,t}^{\perp\prime}(\zeta), z_t^{\perp\prime}(\zeta)\right]'\right] = 0, \tag{6}$$

where $y_{H,t}^{\perp}(d)$, $Y_{H,t}^{\perp}(d)$, $z_t^{\perp}(d)$ are functions of parameters d that depend on the forecasting model chosen, and the true value of d is ζ.

The moments in (5) and (6) can be stacked in a moment function $f(y_{H,t}, Y_{H,t}, z_t, X_{t-1}; b, d)$ with $E[f(y_{H,t}, Y_{H,t}, z_t, X_{t-1}; \beta, \zeta)] = 0$. Let $W_t = [y_{H,t}', \ Y_{H,t}', \ z_t', \ X_{t-1}']'$. The associated GMM objective function is

$$F_T(b, d) = \frac{1}{T}\left(\sum_{t=1}^{T} f(W_t; b, d)\right)' \Phi(b, d) \left(\sum_{t=1}^{T} f(W_t; b, d)\right) \ , \tag{7}$$

where $\Phi(b, d)$ is a positive definite weighting matrix. We assume that Φ is block-diagonal, for instance because the forecast errors are the standard LP or VAR residuals. This ensures that the forecasting step and the structural estimation step are separable for estimation and inference purposes based on a straightforward application of Frisch-Waugh-Lovell. More formally, for inference purposes, we make the assumption,

Assumption 1. *There exists a unique solution, ζ, to the forecasting moments (6), which are linear in X_{t-1}; the associated GMM estimator satisfies $\sqrt{T}\left(\hat{\zeta} - \zeta\right) \xrightarrow{d} \mathcal{N}(0, V_\zeta)$ and for some feasible block-diagonal weighting matrix $\Phi(\beta, \zeta)$ and positive definite V_ζ.*

Under Assumption 1, the Jacobian of (5) with respect to d is zero in expectation at ζ, which, with the structure of $\Phi(\beta, \zeta)$, implies that the asymptotic variance of $\hat{\beta}$ depends only on the asymptotic variance of the sample counterpart of (5), as explained in detail in Appendix C. This means that estimating forecast errors and plugging them into $f_s(\cdot\ , b)$, the part of $f(W_t, b, d)$ corresponding to the structural moments (5), and using

[2] For recent assessments of both methods, see Stock and Watson (2018), Plagborg-Møller and Wolf (2021), or Li et al. (2021).

standard formulas based on just those moments yields a valid asymptotic variance. Henceforth, we therefore take the forecasts as given and focus solely on the structural estimation step. For notational simplicity, we suppress the dependence of the forecast errors on ζ and we use the same "$\perp$" superscript to refer to the sample counterpart of the population forecast errors, since no adjustment is needed in our plug-in approach.

1.2. The SP-IV Estimator

Let $\Phi_s(b,d)$ denote the block in the weighting matrix $\Phi(b,d)$ corresponding to the moments identifying the structural equation, (5). Our baseline estimator uses $\Phi_s(b,d) = I_H \otimes Q^{-1}$, where $Q = E[z_t^\perp z_t^{\perp\prime}]$, to standardize and orthonormalize $z_t^\perp$.[3] In population the solution to (7) identifies β as

$$\begin{aligned}(8) \qquad \beta = &\left(R'(E[Y_{H,t}^\perp z_t^{\perp\prime}]Q^{-1}E[Y_{H,t}^\perp z_t^{\perp\prime}]' \otimes I_H)R\right)^{-1} \\ &\times R' \operatorname{vec}(E[y_{H,t}^\perp z_t^{\perp\prime}]Q^{-1}E[Y_{H,t}^\perp z_t^{\perp\prime}]') \ ,\end{aligned}$$

where $R = I_K \otimes \operatorname{vec}(I_H)$. Let the $H \times T$ matrix $y_H^\perp$, the $HK \times T$ matrix $Y_H^\perp$, and the $N_z \times T$ matrix $Z^\perp$ collect the sample of observations of $y_{H,t}^\perp$, $Y_{H,t}^\perp$, and $z_t^\perp$ respectively. The sample analog of (8) is

$$(9) \qquad \hat\beta \ = \ \left(R'(Y_H^\perp P_{Z^\perp} Y_H^{\perp\prime} \otimes I_H)R\right)^{-1} R' \operatorname{vec}(y_H^\perp P_{Z^\perp} Y_H^{\perp\prime}) \ ,$$

which minimizes (7) with respect to b, using the sample weighting matrix, $I_H \otimes (Z^\perp Z^{\perp\prime}/T)^{-1}$. That minimization problem is equivalent to minimizing $\operatorname{Tr}(u_H^\perp P_{Z^\perp} u_H^{\perp\prime})$, or the sum of squared residuals in the system

$$(10) \qquad y_H^\perp \ = \ (\beta' \otimes I_H)Y_H^\perp + u_H^\perp \ ,$$

after projection on the instruments $z_t^\perp$. Thus, $\hat\beta$ is also the restricted IV estimator in the system of equations in (10), where the only restriction is that β applies at all horizons, as already implied by (1). Because of this formulation, we refer to our framework as **S**ystem **P**rojections on **I**nstrumental **V**ariables (SP-IV), and $\hat\beta$ as the SP-IV estimator.

[3] Efficient GMM uses $\Phi_s(\beta,\zeta) = (\Sigma_{u_H^\perp}^{-1} \otimes Q^{-1})$, where $\Sigma_{u_H^\perp} = E\left[u_{H,t}^\perp(\beta)u_{H,t}^\perp(\beta)'\right]$. Appendix B presents the resulting GLS version of SP-IV, as well as the CUE SP-IV estimator.

The SP-IV estimator has a useful interpretation in terms of the impulse response functions of y_t and Y_t to innovations in the instruments z_t. Consider the following IRF estimates,

$$\hat{\Theta}_Y = \frac{Y_H^\perp Z^{\perp\prime}}{T}\left(\frac{Z^\perp Z^{\perp\prime}}{T}\right)^{-\frac{1}{2}} \quad ; \quad \hat{\Theta}_y = \frac{y_H^\perp Z^{\perp\prime}}{T}\left(\frac{Z^\perp Z^{\perp\prime}}{T}\right)^{-\frac{1}{2}} , \tag{11}$$

which are OLS coefficients from regressing $Y_{H,t}^\perp$ and $y_{H,t}^\perp$ on standardized innovations to the instruments, $\left(Z^\perp Z^{\perp\prime}/T\right)^{-\frac{1}{2}} z_t^\perp$. Using $\hat{\Theta}_y$, construct the $HN_z \times 1$ vector $\hat{\mathbf{\Theta}}_y$ stacking the N_z vectors of IRF coefficients of y_t. Construct the $HN_z \times K$ matrix $\hat{\mathbf{\Theta}}_Y$ similarly stacking $\hat{\Theta}_Y$. Formally,

$$\begin{aligned} \hat{\mathbf{\Theta}}_Y &= ((Z^\perp Z^{\perp\prime}/T)^{-\frac{1}{2}} Z^\perp \otimes I_H/T)\mathbf{Y}_H^\perp \ ; \\ \hat{\mathbf{\Theta}}_y &= ((Z^\perp Z^{\perp\prime}/T)^{-\frac{1}{2}} Z^\perp \otimes I_H/T)\mathbf{y}_H^\perp , \end{aligned} \tag{12}$$

where $\mathbf{y}_H^\perp = \text{vec}\left(y_H^\perp\right)$ is $TH \times 1$ and $\mathbf{Y}_H^\perp = [\text{vec}(Y_{H,1}^\perp), \ldots, \text{vec}(Y_{H,K}^\perp)]$ is $TH \times K$. Then the SP-IV estimator $\hat{\beta}$ in (9) is equivalent to

$$\begin{aligned} \hat{\beta} &= (\mathbf{Y}_H^{\perp\prime}(P_{Z^\perp} \otimes I_H)\mathbf{Y}_H^\perp)^{-1}\mathbf{Y}_H^{\perp\prime}(P_{Z^\perp} \otimes I_H)\mathbf{y}_H^\perp , \\ &= (\hat{\mathbf{\Theta}}_Y'\hat{\mathbf{\Theta}}_Y)^{-1}\hat{\mathbf{\Theta}}_Y'\hat{\mathbf{\Theta}}_y , \end{aligned} \tag{13}$$

so $\hat{\beta}$ is the slope in the OLS regression of $\hat{\mathbf{\Theta}}_y$ on $\hat{\mathbf{\Theta}}_Y$, the coefficients in a regression of the IRFs of y_t and Y_t to z_t, conditional on X_{t-1}.

The expression for $\hat{\beta}$ in (13) suggests a simple two-stage procedure for implementing SP-IV. The first stage consists of estimating IRFs using instruments satisfying the exogeneity conditions, which are frequently estimated objects in empirical macroeconomics. In the second stage, given such a set of IRF estimates, the SP-IV estimator is obtained by regressing the IRF of the outcome variable, y_t, on the IRFs of the endogenous variables, Y_t. To theoretically justify the moment conditions in (4), it will often be natural to choose instruments leading to impulse responses to interpretable economic shocks, such as monetary policy shocks, government spending shocks, etc. For the Phillips curve example in (2), the first stage estimates IRFs of inflation π_t and the slack measure gap_t to a monetary policy shock (or other aggregate demand shocks orthogonal to

the cost-push term, u_t). In the second stage, the IRF of π_t is regressed on the IRF of gap_t as well as the IRFs of lagged and expected future inflation, π_{t-1} and π^e_{t+1}. The latter can be obtained simply by lagging and leading the IRF of π_t by one horizon. Appendix A gives practical details on implementation using LPs or VARs.

A large literature studies the identification of economic shocks presenting potential instruments for SP-IV; see Ramey (2016) or Kilian and Lütkepohl (2017) for surveys. However, any valid strategy for identifying structural IRFs based on LPs or VARs can be used in conjunction with SP-IV provided the underlying shocks satisfy the exogeneity conditions (5). SP-IV recovers the coefficients that best fit the structural economic relationship between IRFs of the variables in that relationship to the shocks chosen by the econometrician. Technically, SP-IV only requires IRFs to an identified rotation of economic shocks that satisfy the exogeneity conditions. In other words, the shocks and their associated IRFs need not necessarily be separately identified. In practice, it is also possible to perform SP-IV with a subset of horizons rather than all $h = 0, \ldots, H-1$, as we discuss further below.

1.3. SP-IV versus 2SLS Implementations of SE-IV

The standard SE-IV approach for identifying β in (1) with $z_t, ..., z_{t-H+1}$ as instruments exploits the HN_z orthogonality conditions

$$E[u_t z_{t-h}] = 0 \quad ; \quad h = 0, \ldots, H-1 \ . \tag{14}$$

Practitioners typically implement these conditions using the 2SLS estimator, in which case the first stage consists of regressing the endogenous variables Y_t on the lag sequence $z_t, ..., z_{t-H+1}$, and the second stage consists of regressing y_t on the predicted values. When z_t consists of measures of economic shocks, the first stage implicitly estimates the IRF coefficients of Y_t to the shocks z_t using a DL model. Barnichon and Mesters (2020) observe that, after similarly estimating the IRF of y_t, the 2SLS estimates equal the estimates from OLS regression of the IRF of y_t on the IRFs of Y_t. The 2SLS estimator with a DL of shocks as instruments, therefore,

can – like SP-IV – be interpreted in terms of a regression with IRFs. In 2SLS, the regression uses IRFs estimated by single-equation DL models, i.e., regressions of y_t and Y_t on $z_t, ..., z_{t-H+1}$ without additional controls. In contrast, in SP-IV the IRFs can be obtained from LPs with controls X_{t-1} or from VARs with y_t, Y_t and other variables in X_t as endogenous variables. In the literature, IRFs are typically identified from such LPs or VARs with additional controls, not DL models, since plausible unconditionally exogenous instruments are generally hard to find. One advantage of SP-IV is, therefore, that it estimates structural relationships across IRFs as they are estimated in practice. Another advantage is that SP-IV greatly expands the options for identification. Whereas the IRFs for 2SLS rely on the availability of external measures of economic shocks, the IRFs for SP-IV can also be identified by internal instruments generated from recursivity assumptions, or any other covariance restrictions that do not involve external variables.

Depending on the DGP, the ability to accommodate controls yields three further potential advantages of SP-IV.

1. Weaker Exogeneity Requirements for z_t The first advantage is that, with suitable predetermined controls X_{t-1}, practical identification arguments can be based on weaker assumptions regarding z_t. To see this more clearly, we adopt the conventional impulse-propagation paradigm of representing y_t and Y_t in terms of linear combinations of current and past realizations of 'structural shocks', ϵ_t, where $E[\epsilon_t] = 0$, $E[\epsilon_t \epsilon_t'] = I_{dim(\epsilon)}$ and $E[\epsilon_t \epsilon_s'] = 0$ for $s \neq t$.[4] Such a representation is consistent with a large class of DGPs, including all stationary SVAR models (from a reduced-form perspective) and all discrete-time linearized DSGE models (from a structural perspective); see, e.g., Plagborg-Møller and Wolf (2021). Given the representations for y_t and Y_t as linear combinations of the history of ϵ_t, (1) implies that the error term u_t is generally also a linear combination

[4]This paradigm is common in the literature, e.g., Ramey (2016), Stock and Watson (2018), Plagborg-Møller and Wolf (2021) or Plagborg-Møller and Wolf (2022) for recent examples.

of current and past structural shocks:

$$u_t = \mu_0'\epsilon_t + \mu_1'\epsilon_{t-1} + \mu_2'\epsilon_{t-2} + \ldots, \tag{15}$$

i.e., u_t is an MA(∞) in the structural shocks.

Given the representation of the error term u_t in (15) and stationarity, the necessary orthogonality condition for SE-IV in (14) can be restated as $\sum_{l=0}^{\infty} \mu_l' E[\epsilon_{t+h-l} z_t'] = 0$. Without any controls, substituting (15) into the orthogonality condition for SP-IV in (4) yields exactly the same orthogonality condition. Although it is theoretically possible for nonzero terms in the summation of this condition to cancel each other out, practical identification approaches will almost always rely on the sufficient conditions that each individual term in the summation is zero since, otherwise, a very particular relationship must be assumed to hold among the nonzero μ_l coefficients. With suitably chosen controls X_{t-1}, these sufficient conditions are weaker for SP-IV than for SE-IV:

Proposition 1. *Suppose* (15) *and stationarity hold; the exogeneity condition for SE-IV with lags of* z_t *in* (14) *holds when*

$$\mu_l' E[\epsilon_{t+h-l} z_t'] = 0 \quad ; \quad l = 0, \ldots, \infty \quad ; \quad h = 0, \ldots, H-1. \tag{16}$$

Suppose (15) *holds and* X_{t-1} *spans past shocks included in* u_t *such that* $u_t^{\perp} = \mu_0'\epsilon_t$*; the exogeneity condition for SP-IV in* (4) *holds when*

$$\mu_l' E[\epsilon_{t+h-l} z_t'] = 0 \quad ; \quad l = 0, \ldots, h \quad ; \quad h = 0, \ldots, H-1. \tag{17}$$

Proof. The SE-IV result follows from substituting (15) in (14) and stationarity. The SP-IV result follows similarly after conditioning on X_{t-1}. □

Analogous to Stock and Watson (2018), we denote the sufficient conditions in (16) with $l > h$ as *lag exogeneity*, with $l = h$ as *contemporeneous exogeneity*, and with $l < h$ as *lead exogeneity*. The SE-IV exogeneity condition holds when all three forms of exogeneity hold. In contrast, the SP-IV exogeneity condition is implied by only contemporaneous and lead exogeneity, since by assumption conditioning on X_{t-1} eliminates the

influence of all past realizations of ϵ_t on $u_t^{\perp}(h)$. With a suitable set of predictors, the exogeneity conditions on z_t are thus substantially weaker.[5]

Furthermore, there are alternative intuitive sufficient conditions for the SP-IV exogeneity conditions in (4). For instance, if X_{t-1} instead spans any past shocks entering z_t, so that $z_t^{\perp}$ is a function of only contemporaneous shocks other than those entering u_t, then the conditions are satisfied. Indeed, various combinations of conditions like this one and that with respect to $u_t^{\perp}$ in Proposition 1 will imply (4).

As a concrete example of how controls can matter for instrument validity, consider again the Phillips Curve in (2). As instruments, Barnichon and Mesters (2020) use a DL of Romer and Romer's (2004) measure of monetary policy surprises, z_t^{RR}, which are the residuals in a regression of the intended funds rate change at FOMC meetings on the current rate and Greenbook forecasts of output growth and inflation. Assume no measurement error and that the error term in (2) is just an exogenous cost-push shock following $u_t = \rho_u u_{t-1} + \upsilon_t$, with $0 \leq \rho_u < 1$, and υ_t is white noise. Unless $\rho_u = 0$, u_t depends on υ_t, and on all lags $\upsilon_{t-1}, \upsilon_{t-2}, \ldots$ If z_t^{RR} is uncorrelated with υ_t, its leads up to $H-1$, and all of its lags, then $z_t^{RR}, ..., z_{t-H+1}^{RR}$ satisfy the sufficient conditions (16) for estimation of the Phillips Curve. Suppose, however, that the regression generating z_t^{RR} is misspecified by omitting one or more lags of inflation. In that case, z_t^{RR} generally still depends on lags of υ_t, and the lag exogeneity requirement is not satisfied. However, by including lags of inflation amongst predictors X_{t-1}, the exogeneity requirements for SP-IV remain satisfied as long as contemporaneous and lead exogeneity hold. We return to this example later in the simulations of Section 3.

The identification results in Proposition 1 mirror those in Stock and Watson (2018) showing that the inclusion of suitable lagged controls in LP-IV, or equivalently 'invertibility' in SVARs, avoids lag exogeneity requirements when estimating structural impulse responses. Here though,

[5] If $u_{t+h}^{\perp} = \sum_{j=0}^{h} \mu_j' \epsilon_{t+h-j}$ for $h = 0, ..., H-1$, the necessary condition for SP-IV is $\sum_{l=0}^{h} \mu_l' E[\epsilon_{t+h-l} z_t^{\perp\prime}] = 0$, compared to the necessary condition $\sum_{l=0}^{\infty} \mu_l' E[\epsilon_{t+h-l} z_t'] = 0$ for SE-IV. With nonzero terms in the summations, it is strictly speaking possible that the necessary condition holds for SE-IV but not for SP-IV. However, we are unaware of any application in the literature with identification arguments that would imply nonzero terms in either summation.

the requirements are in principle weaker than those needed to estimate dynamic causal effects using LPs of SVARs with lag endogenous instruments, since conditioning on X_{t-1} must remove only the influence of lagged shocks that are included in (15) – i.e. shocks in ϵ_t with corresponding nonzero elements in μ_l – and not necessarily that of lags of *all* shocks ϵ_t driving y_t and Y_t.[6]

2. Efficiency Gains Conditioning on predictors X_{t-1} can also lead to asymptotic efficiency gains relative to 2SLS with DL instruments. Whether SP-IV improves efficiency depends on the DGP for u_t and the informativeness of the predictors X_{t-1}. Intuitively, the SP-IV estimator is more efficient than 2SLS if the variances of the forecast errors $u_t^{\perp}(h)$ at $h = 0, ..., H-1$ are small relative to the variance of the error term u_t. Let $\Sigma_{u_H^{\perp}} = \text{var}(u_{H,t}^{\perp})$ and $\Sigma_{u_H} = \text{var}(u_{H,t})$.

Proposition 2. *Suppose z_t is i.i.d. and independent of u_t; then*

(i) If u_t is i.i.d., or if X_{t-1} is empty or is otherwise uninformative for $u_t, \ldots, u_{t+H-1}$, SP-IV is asymptotically as efficient as 2SLS.

(ii) $\hat{\beta}$ is asymptotically more efficient than $\hat{\beta}_{2SLS}$ if $\text{maxeval}\big(\Sigma_{u_H}\big) > \text{maxeval}\left(\Sigma_{u_H^{\perp}}\right)$.

Proof. See Appendix D. □

When u_t is i.i.d., the errors in both estimators are identical in population since X_{t-1} does not predict $u_t, \ldots u_{t+H-1}$ and forecast errors do not accumulate over $h = 0, \ldots, H-1$. More generally, if X_{t-1} is uninformative for $u_t, \ldots u_{t+H-1}$, the forecast error variance of $u_t(h)^{\perp}$ accumulates in exactly the same way as autocovariance terms in the long-run variance of the 2SLS estimator; intuitively, if the estimators use equivalent moments (under stationarity) and weighting, their asymptotic variances must be identical. Part (ii) states that SP-IV can be asymptotically more efficient than 2SLS if X_{t-1} has sufficient predictive power for $u_t, \ldots u_{t+H-1}$. If u_t

[6] In the Phillips curve example, other demand shocks could still contaminate z_t^{RR} after conditioning on X_{t-1}, and the IRFs identified with z_t^{RR} in VARs or LPs with X_{t-1} as controls may therefore not represent the causal effects of monetary policy shocks; nevertheless, z_t^{RR} remains a valid instrument for SP-IV with controls X_{t-1} as long as X_{t-1} eliminates the influence of $v_{t-1}, v_{t-2}, \ldots$

follows an AR(1), for example, and X_{t-1} spans all past shocks, applying this formula shows that SP-IV is more efficient than 2SLS for any $H > 1$ and $|\rho| > 0$. Intuitively, efficiency gains from SP-IV are more likely when u_t is more persistent such that X_{t-1} explains a larger fraction of the variance of $u_t, \dots u_{t+H-1}$. Since the predictive power of X_{t-1} diminishes as the forecast horizon grows larger, relative efficiency of SP-IV requires that the maximum forecast horizon, H, is not too large.

3. Stronger Identification The ability to condition on X_{t-1} in SP-IV can also improve the effective strength of the instruments, as measured by the concentration parameter; see, e.g., Stock and Yogo (2005). Weak instruments lead to bias in IV estimators and make conventional inference methods invalid. In many time series applications, instruments are weak, while the endogenous variables can be highly persistent and thus predictable. Let ω_t be the error term in the first stage of 2SLS with variance σ^2_ω. Denote the $H \times 1$ vector of errors in the SP-IV first stage regression of $Y_{H,t}$ on z_t (but no additional predictors) as $v_{H,t}$, with covariance Σ_{v_H}. Similarly, denote the errors in the first-stage SP-IV regression of $Y_{H,t}$ on z_t and the additional predictors X_{t-1} by $v^\perp_{H,t}$, with covariance $\Sigma_{v^\perp_H}$. Finally, consider the case where z_t is strictly exogenous (i.e., lead/lag and contemporaneously exogenous) since otherwise, the concentration parameter for 2SLS is not well-defined.

Proposition 3. *Assume $K = 1$ and z_t is i.i.d. and satisfies (16):*

(i) Unless $Y^\perp_t(h) = Y_{t+h}$ for $h = 0, ..., H-1$, such that X_{t-1} are irrelevant predictors, the concentration parameter for SP-IV with controls X_{t-1} is larger than for SP-IV without controls;

(ii) If $\mathrm{Tr}(\Sigma_{v^\perp_H})/H < \sigma^2_\omega$, the concentration parameter for SP-IV is larger than that for 2SLS.

Proof. See Appendix E. □

Part (i) in Proposition 3 states that when the predictors have explanatory power for the endogenous regressors, their inclusion in SP-IV increases the effective strength of the instruments, as measured by the concentration parameter, and conditioning on X_{t-1} therefore decreases bias.

Part (ii) states that the effective instrument strength can also increase relative to 2SLS, depending on the persistence and predictability of the errors, as well as on H. As the predictability of the endogenous variables diminishes with the forecast horizon H, the advantage of conditioning on X_{t-1} can be outweighed by the recency of z_t for Y_t in 2SLS. With multiple endogenous regressors ($K > 1$), instrument strength depends on the entire eigenstructure of the first stage parameters (and that of $\Sigma_{v_H^\perp}$), making a fully general result hard to obtain analytically. Intuitively, however, conditioning on X_{t-1} should still strengthen the instruments when X_{t-1} has explanatory power and $K > 1$.

Each of the three potential advantages of SP-IV above derives from the ability to include lagged controls when estimating the first-stage IRFs, and they will be demonstrated by the simulation evidence in Section 3 below. As we illustrate with further simulation results in Online Appendix II.4, it is *not* possible to replicate the same advantages by incorporating the controls X_{t-1} into 2SLS implementations of the conventional SE-IV setting. First, adding X_{t-1} as additional regressors in both stages of 2SLS with DL instruments weakens identification. As an extreme case, suppose conditioning on X_{t-1} eliminates the influence of all past realizations of the structural shocks ϵ_t on Y_t and z_t. Including X_{t-1} as additional regressors in 2SLS then implies that only the contemporaneous instruments z_t remain relevant. Since X_{t-1} also removes the influence of lags of z_t, these lags are completely irrelevant as instruments after including X_{t-1} as controls. Moreover, when $N_z < K$ the model is under-identified without identifying information from the lags of z_t.

Second, it is generally also not possible to circumvent the lag exogeneity requirement of 2SLS by first projecting z_t on X_{t-1} and using the residuals, $z_t^\perp, ..., z_{t-H+1}^\perp$ as the instrumental variables in 2SLS. This is the implicit procedure, for example, when a shock is first identified in a VAR or LPs with X_{t-1} as controls, and a DL of that shock is then used as instruments in 2SLS. The reason is that u_t must still be orthogonal to all lags of the identified shock, which is not necessarily the case. Even when such a weaker form of lag exogeneity is plausible, this approach does not have the

other advantages of SP-IV when X_{t-1} has substantial predictive power.[7]

Finally, adding X_{t-H} to both stages of a 2SLS regression of y_t on Y_t with $z_{t-1}, ...z_{t-H+1}$ as instruments also does not replicate the advantages of SP-IV. This approach is based on instrumenting more distant $H-1$-step ahead forecast errors and still requires lag exogeneity at horizons smaller than H. Moreover, by instrumenting only more distant $H-1$-step-ahead forecast errors, this approach does not increase identification strength to the same extent as SP-IV with X_{t-1} as controls.

We conclude this section with some practical considerations for choosing the controls in empirical applications. First, the choice of controls matters in cases when the instruments z_t are suspected to be lag endogenous. In that case, conditioning on X_{t-1} must remove the influence of all past structural shocks that are not excluded from the structural MA(∞) representation of the error term u_t in (15) from $u_t^\perp$, $z_t^\perp$, or both. So if u_{t+h} is not i.i.d., but, for example in the Phillips curve application, depends on the realizations of cost-push shocks before period t, then X_{t-1} must be such that the period t $h+1$-step ahead conditional forecast error, $u_t^\perp(h)$, is orthogonal to the cost-push shocks prior to period t, the orthogonalized instrument $z_t^\perp$ is orthogonal to cost-push shocks prior to period t, or both. When u_{t+h} is also influenced by other structural shocks, for example, because of mismeasurement of inflation expectations or the output gap, then conditioning on X_{t-1} must generally remove the influence of the entire history of those other confounding shocks as well. Researchers who want to remain entirely agnostic about which structural shocks prior to t influence u_{t+h} can aim for a set of controls X_{t-1} that is informationally equivalent to the (infinite) history of *all* macroeconomic shocks that drive y_t and Y_t. The latter provides the greatest amount of 'exogeneity insurance' – as Ramey (2016) calls it in the context of structural impulse response estimation. It is also a testable assumption, and various tests are available in the literature. As mentioned before, the controls X_{t-1} can also include lags of z_t itself, such that the first-stage IRFs can be esti-

[7] See also Lloyd and Manuel (2023) for further discussion of why residualizing instruments – but not regressors – to controls is not advisable.

mated using 'lag-augmented' local projections as discussed in Montiel Olea and Plagborg-Møller (2021) or internal instrument VARs as discussed in Plagborg-Møller and Wolf (2021).

Second, even when the instruments z_t are strictly exogenous and the controls X_{t-1} are not strictly necessary for valid identification, the choice of controls still matters for realizing the efficiency and/or instrument strength improvements of SP-IV relative to 2SLS. Intuitively, these benefits are maximized when the forecast errors have the smallest possible variance, and this will be the case when X_{t-1} contains all relevant predictive information, i.e. when X_{t-1} is informationally equivalent to the (infinite) history of all structural shocks that drive y_t and Y_t. Each of the above two roles for X_{t-1} (exogeneity insurance and efficiency/ stronger identification) point to including all relevant predictive information for the endogenous variables in X_{t-1}. In this sense, the considerations for choosing the controls are not different from those when estimating structural IRFs, see Stock and Watson (2018). Of course, choosing larger X_{t-1} in practice must be weighed against small sample issues, and taking maximum advantage could involve lagged factors from factor models, shrinkage, or Bayesian methods. We leave these refinements for future work.

1.4. Consistency of the SP-IV Estimator

Consider the following high-level assumptions on covariances:

Assumption 2. *The following probability limits and rank condition hold:*

(2.a) $Z^{\perp}Z^{\perp\prime}/T \xrightarrow{p} E[z_t^{\perp}z_t^{\perp\prime}] = Q$, *where Q is positive definite,*

(2.b) $Y_H^{\perp}Z^{\perp\prime}/T \xrightarrow{p} E[Y_{H,t}^{\perp}z_t^{\perp\prime}] = \Theta_Y Q^{\frac{1}{2}}$, *a real $HK \times N_z$ matrix,*

(2.c) $Z^{\perp}u_H^{\perp\prime}/T \xrightarrow{p} E[z_t^{\perp}u_{H,t}^{\perp\prime}] = 0$,

(2.d) $R'(\Theta_Y\Theta_Y' \otimes I_H)R$ *is a fixed matrix with full rank.*

The convergence in probability in 2.a-2.c holds under standard primitive conditions and laws of large numbers. Condition 2.a ensures linear independence of the instruments and consistency of the sample weighting matrix. Condition 2.b states that the covariance between $Y_H^{\perp}$ and $Z^{\perp}$ is

consistently estimated. The population covariance $\Theta_Y Q^{\frac{1}{2}}$ is a rotation of Θ_Y, a matrix containing the impulse response coefficients of $Y_t^\perp$ to $z_t^\perp$, after standardization. Condition 2.c is the exogeneity condition. Finally, the rank condition 2.d is sufficient for the existence of a unique solution to the moment conditions (5) and ensures that the denominator of the closed form solution (8) is full rank, with the definition of Θ_Y implying that the instruments are relevant. 2.b and 2.d jointly imply that the instruments are strong, an assumption we relax in Section 2.

Assumption 2 resembles the usual (strong) IV assumptions; see, for instance, Stock and Yogo (2005). Note that condition 2.d does not require that there are at least as many instruments as endogenous regressors, $N_z \geq K$. Since $\operatorname{rank}(R'(\Theta_Y\Theta_Y' \otimes I_H)R) = \min\{K, H \operatorname{rank}(\Theta_Y\Theta_Y')\}$, the order condition is $HN_z \geq K$, since there are HN_z moment conditions in (5). Adding leads of y_t and Y_t makes up for $N_z < K$ just as adding lags of z_t as instruments does for SE-IV.

Proposition 4 establishes consistency of the SP-IV estimator in (9).

Proposition 4. *Under Assumptions 1 and 2,* $\hat{\beta} \xrightarrow{p} \beta$.

Proof. Both terms in (9) converge by the stated assumptions, and the result follows from the continuous mapping theorem. □

2. Inference for SP-IV

2.1. Inference under Strong Instruments

When the instruments are strong, under the conditions in Assumptions 1 and 2, inference for SP-IV can proceed analogously to standard 2SLS. With a further high-level assumption, the limiting distribution of $\hat{\beta}$ follows:

Assumption 3. $T^{-1/2} \operatorname{vec}(Z^\perp u_H^{\perp\prime}) \xrightarrow{d} N(0, (\Sigma_{u_H^\perp} \otimes Q))$, *where* $\Sigma_{u_H^\perp}$ *is full rank.*

Proposition 5. *Under Assumptions 1-3,*

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta) \ , \tag{18}$$

where $V_\beta = (R'(\Theta_Y\Theta_Y' \otimes I_H)R)^{-1} R' \left(\Theta_Y\Theta_Y' \otimes \Sigma_{u_H^\perp}\right) R \left(R'(\Theta_Y\Theta_Y' \otimes I_H)R\right)^{-1}$.

Proof. The result is immediate after rearranging (9), from Proposition 4, the stated assumptions, and the continuous mapping theorem. □

V_β can be estimated by replacing $\Sigma_{u_H^\perp}$ with a consistent estimate, and $\Theta_Y \Theta_Y'$ with $Y_H^\perp P_{Z^\perp} Y_H^{\perp\prime}$. Inference can be based on standard Wald tests. A natural consistent estimator is

$$\hat{\Sigma}_{u_H^\perp} = \hat{u}_H^\perp \hat{u}_H^{\perp\prime}/(T - N_x - K), \tag{19}$$

since, as noted following Assumption 1, estimation error in the forecast errors does not impact their asymptotic variances. Including adequate lags in X_{t-1} (which can include lags of z_t) obviates the need for an autocorrelation robust estimate by eliminating autocorrelation in $z_t^\perp$. Any mechanical correlation between $u_t^\perp(0)$ and $u_{t-h}^\perp(h)$, say, drops out of $\text{var}(u_{H,t}^\perp \otimes z_t^\perp)$, since when $z_t^\perp$ is serially uncorrelated, so too is $u_{H,t}^\perp \otimes z_t^\perp$.[8] This is not the case for 2SLS, which generally requires autocorrelation-robust methods due to mechanical autocorrelation in the overlapping lag sequence of z_t. The same logic also prevents strong dependence in the forecast errors at long horizons from causing spurious regression problems when estimating first-stage or reduced-form IRFs.

2.2. A Test for Weak Instruments

In many applications, the available instruments may be weak. If so, Wald inference will be invalid, leading to empirical rejection rates that generally exceed nominal levels. In the Online Appendix, we derive a bias-based test of instrument strength for SP-IV that is analogous to the popular Stock and Yogo (2005) bias-based test of weak instruments for standard 2SLS. We consider a Nagar approximation of the bias under weak instrument asymptotics, as in Montiel-Olea and Pflueger (2013) and Lewis and Mertens (2022). Like Stock and Yogo (2005) and Lewis and Mertens (2022), we use a weighted ℓ_2-norm of the bias to accommodate multiple endogenous regressors ($K>1$). Weak instruments are defined as those for which the bias in $\hat{\beta}$ is at least τ percent of a worst-case benchmark under

[8]The argument is analogous to that of Montiel Olea and Plagborg-Møller (2021) for LP with instrumental variables.

weak instrument asymptotics. The test statistic is similar to that of Cragg and Donald (1993), and the test rejects the null hypothesis of weak instruments when the statistic exceeds the level-α critical value of a bounding distribution. The test nests the Stock and Yogo (2005) test when $H = 1$.

2.3. Weak Instrument Robust Inference for SP-IV

We describe two robust test statistics for SP-IV with local projections. Appendix A describes the implementation of the tests when using a VAR.

AR Statistic The "S-statistic" of Stock and Wright (2000) extends the AR statistic to the GMM setting. For SP-IV, the statistic and its limiting distribution under the null hypothesis are

$$AR(b) = (T - d_{AR}) \operatorname{Tr}\left(u_H^{\perp}(b) P_{Z^{\perp}} u_H^{\perp}(b)' \left(u_H^{\perp}(b) M_{Z^{\perp}} u_H^{\perp}(b)'\right)^{-1}\right), \tag{20}$$

$$AR(\beta) \xrightarrow{d} \chi^2_{HN_z} \ ,$$

where $M_{Z^{\perp}} = I_T - P_{Z^{\perp}}$ is the residualizing matrix and $d_{AR} = N_z + N_x$ is a degrees of freedom correction. Rather than the moment covariance matrix, we use the normalizing matrix typically used with the AR statistic, asymptotically equivalent under the null hypothesis. Note that estimation error in the first stage does not affect inference based on the AR statistic since it does not impact the asymptotic variance of the $u_{H,t}^{\perp} \otimes z_t^{\perp}$, for the same reasons given under Assumption 1.

KLM Statistic The AR statistic can have poor power when there are over-identifying restrictions. This occurs when $HN_z > K$, i.e., when the number of IRF coefficients exceeds the number of endogenous regressors. As this will often be the case, we consider the Kleibergen (2005) KLM statistic, which can improve power (Andrews et al. 2019).

Following Kleibergen (2005),

$$
\begin{aligned}
(21) \qquad K(b) &= (T - d_K)\,\mathrm{vec}\left(\Xi^{-1}u_H^{\perp}(b)\check{Y}_H'\right)' R \\
&\quad \times \left(R'(\check{Y}_H\check{Y}_H' \otimes \Xi^{-1}u_H^{\perp}(b)u_H^{\perp\prime}(b)\Xi^{-1})R\right)^{-1} \\
&\quad \times R'\,\mathrm{vec}\left(\Xi^{-1}u_H^{\perp}(b)\check{Y}_H'\right)' \ , \\
K(\beta) &\xrightarrow{d} \chi^2_K \ ,
\end{aligned}
$$

where $\check{Y}_H = Y_H^{\perp}P_{Z^{\perp}} - \check{v}_H^{\perp}\check{u}_H^{\perp\prime}(b)\left(\check{u}_H^{\perp}(b)\check{u}_H^{\perp\prime}(b)\right)^{-1}u_H^{\perp}(b)P_{Z^{\perp}}$, $\Xi = u_H^{\perp}(b)M_{Z^{\perp}}u_H^{\perp\prime}(b)$, $\check{v}_H^{\perp} = v_H^{\perp}M_{Z^{\perp}}$, $\check{u}_H^{\perp}(b) = u_H^{\perp}(b)M_{Z^{\perp}}$, and $d_K = N_z + N_x$ is a degrees of freedom correction. Intuitively, instead of the covariance of $u_H^{\perp}$ and $(Z^{\perp}Z^{\perp\prime})^{-1/2}Z^{\perp}$, the numerator of the KLM statistic features the covariance of $u_H^{\perp}$ and the projection of a transformation of $Y_H^{\perp}$ on $(Z^{\perp}Z^{\perp\prime})^{-1/2}Z^{\perp}$. Our formulation differs from Kleibergen (2005) only by the replacement of $u_H^{\perp}$ and $v_H^{\perp}$ with $\check{u}_H^{\perp}$ and $\check{v}_H^{\perp}$. This choice is consistent with the IV statistic in Kleibergen (2002) and asymptotically equivalent to the form in Kleibergen (2005) under the null.

3. Performance of SP-IV in Model Simulations

In this section, we demonstrate the performance improvements offered by SP-IV in simulations, supporting our theoretical claims. The objective in all simulations is to estimate the parameters of the Phillips Curve in (2) using data generated from the macroeconomic model of Smets and Wouters (2007) (SW).[9] The Phillips Curve in (2) is one of the equations in the SW model within a system of fourteen simultaneous equations for the dynamics of key macroeconomic aggregates. An important feature of the estimated SW model is that the shocks underlying the error term u_t in the Phillips curve explain a very large fraction of the variance of inflation. This means that, in realistic sample sizes, the weak instrument problem is generally severe due to the small role played by other shocks. Moreover, the error term u_t is persistent, as are most of the macro aggregates generated by the model. Both features make the estimation of

[9]The data is generated from the SW model using the Dynare replication code kindly provided by Johannes Pfeifer at https://sites.google.com/site/pfeiferecon/dynare.

the Phillips curve parameters challenging. Conventional SE-IV methods tend to perform poorly, and our simulation setup is, therefore, an ideal laboratory to evaluate the potential improvements offered by SP-IV.

As mentioned in the introduction, using a sequence of lagged endogenous variables as instruments – as in Galí and Gertler (1999) and the subsequent literature – is not valid for identification in this setting. In the SW model, the error term in (2) is the ARMA(1,1) process

$$u_t = \rho_u u_{t-1} + \epsilon_t^p - \mu_p \epsilon_{t-1}^p \ , \ \rho_u = 0.99, \ \mu_p = 0.83 \tag{22}$$

where ϵ_t^p is an i.i.d. normally distributed price markup shock.[10] Inverting the autoregressive term in (22) yields $u_t = \epsilon_t^p + \rho_u(1-\mu_p)\epsilon_{t-1}^p + \rho_u(\rho_u - \mu_p)\epsilon_{t-2}^p + \rho_u^2(\rho_u - \mu_p)\epsilon_{t-3}^p + \ldots$, showing that the error term u_t generally depends on the entire history of price markup shocks $\epsilon_t^p, \epsilon_{t-1}^p, \epsilon_{t-2}^p, \ldots$ The period t values of the endogenous model variables are functions of all current and lagged values of a 7×1 vector ϵ_t, including ϵ_t^p. Lagged values of these endogenous variables thus violate the lag exogeneity requirement.

Because lagged endogenous variables are not valid instruments, we consider a measure of the monetary policy shock as z_t, as in Barnichon and Mesters (2020). We present two sets of simulations. In the first, we use a measure of monetary policy shocks that violates the lag exogeneity requirement in an arguably realistic manner to illustrate that the SP-IV estimator – unlike the 2SLS estimator – remains consistent. In the second, we use the true model monetary policy shock as the instrument to level the playing field across estimators and compare the small sample performance of 2SLS and SP-IV when both are consistent.

Inflation expectations π_{t+1}^e are treated as unobserved and are replaced in (2) by realized future inflation π_{t+1}, as is common in the literature when expectations appear in structural equations. Under rational expectations – as assumed in the SW model – the resulting measurement error depends only on future realizations of the model shocks, which does not create any additional endogeneity problems given that the instruments used in all simulations satisfy lead exogeneity.

[10] We assume that the econometrician cannot exploit the ARMA(1,1) error structure in (22).

We do not assume that the econometrician possesses a set of controls spanning the full history of model shocks. Instead, we use a realistic set of controls consisting of seven endogenous model variables: the short-term interest rate, inflation, marginal cost, output, consumption, investment, and the real wage. In the simulations, we consider both LP and VAR implementations of SP-IV that include four lags of these endogenous model variables in the control set, X_{t-1}. Note that the SW model does not permit a finite-order structural VAR representation in the seven endogenous model variables listed above. As shown in Plagborg-Møller and Wolf (2021), this implies that the VAR-based estimates of the IRFs are (asymptotically) biased for horizons beyond the lag length of the VAR, whereas the LP-based IRFs remain consistent at all horizons as long as the necessary exogeneity requirements are satisfied. This disadvantage of VAR-based IRFs must be weighed against the efficiency gains relative to LP-based IRFs, see Li et al. (2021). We consider both LP and VAR implementations of SP-IV in the simulations, as it is not clear *ex ante* how the different bias-variance properties of the IRF estimators translate to the SP-IV estimators.

We also report key insights from additional simulations for specifications with multiple demand shocks as instruments, for the generalized (or efficient GMM) versions of the SP-IV estimators, and for alternative specifications for 2SLS that incorporate controls. The full results for these additional simulations are available in the Online Appendix.

3.1. Simulations with Violations of Lag Exogeneity

Our first set of simulations demonstrates how SP-IV can help ensure exogeneity by conditioning on lagged macroeconomic variables. We are motivated by the identification of the Phillips Curve, for example, with monetary policy shock measures like those constructed by Romer and Romer (2004), or based on high-frequency changes in Fed Funds futures as in Kuttner (2001). A practical concern with such measures is that, despite careful construction, they may still contain a meaningful predictable component (Barakchian and Crowe 2013; Bauer and Swanson 2022; Cieslak 2018; Coibion 2012; Miranda-Agrippino and Ricco 2021; Ramey 2016).

Consequently, researchers identifying monetary IRFs using these measures typically include various lagged macro variables as controls in their models. However, when the same measures are used as instruments to estimate structural equations using 2SLS – as in Barnichon and Mesters (2020) for example – estimation proceeds without controls.

To illustrate the potential implications of excluding controls, we simulate "Romer and Romer (2004) instruments" that consist of the true monetary policy shocks in the SW model, augmented with a linear function of inflation over the past four quarters. We calibrate the coefficients on lagged inflation by regressing the actual Romer and Romer (2004) measures on four lags of the log change in the GDP deflator (the inflation measure used to estimate the SW model) over the 1969-2004 sample. The resulting instruments have non-zero covariances with lagged inflation that match the U.S. data (with an R^2 of 0.08) and therefore violate the lag exogeneity requirement. However, by construction, the simulated instruments are exogenous conditional on X_{t-1} since the relevant lags of inflation are included among the variables in the control set.

The left panel in Table 1 reports mean estimates of $\beta = [\gamma_b, \gamma_f, \lambda]'$ across 5000 Monte Carlo samples. We consider specifications with horizons of $H = 8$ and $H = 20$ quarters. To minimize small-sample features and focus on the violation of the exogeneity requirements, Table 1 considers a long sample $T = 5000$. The true model parameters are shown in the first row, with OLS estimates in the second. The remaining rows report results for 2SLS with H lags of the monetary policy instrument, SP-IV based on LP without controls (SP-IV LP), and LP and VAR implementations of SP-IV (LP-C and VAR) that include controls X_{t-1}.

Unsurprisingly, the OLS estimates are severely biased because of endogeneity, pointing incorrectly to a completely flat Phillips curve. Because of the violation of lag exogeneity, the estimates based on 2SLS with DL instruments are also strongly biased. The average estimate of λ even has the wrong sign for both $H = 8$ and $H = 20$. The next row shows the SP-IV estimator without controls X_{t-1}; it is also biased because, like 2SLS, it requires lag exogeneity to hold. The bias is almost identical to that of 2SLS since, in this case, SP-IV and 2SLS exploit very similar sample

TABLE 1: RESULTS WITH LAG ENDOGENOUS INSTRUMENT, $T = 5000$

	Mean Estimates		
	γ_b	γ_f	λ
True Value	0.15	0.85	0.05
OLS	0.48	0.48	0.00
$H = 8$			
2SLS	0.27	0.58	-0.09
SP-IV LP	0.26	0.60	-0.08
SP-IV LP-C	0.16	0.84	0.05
SP-IV VAR	0.12	0.83	0.09
$H = 20$			
2SLS	0.24	0.76	-0.02
SP-IV LP	0.24	0.75	-0.02
SP-IV LP-C	0.23	0.81	0.02
SP-IV VAR	0.17	0.83	0.05

Empirical Size of Nominal 5% Tests		
	$H = 8$	$H = 20$
Wald 2SLS	55.0	96.0
Wald SP-IV LP	61.2	96.0
Wald SP-IV LP-C	9.3	34.9
Wald SP-IV VAR	5.5	13.3
AR SE-IV	72.5	72.3
AR SP-IV LP	67.6	54.4
AR SP-IV LP-C	4.6	5.8
AR SP-IV VAR	4.9	4.5
KLM SE-IV	82.5	85.9
KLM SP-IV LP	81.5	72.4
KLM SP-IV LP-C	5.2	5.5
KLM SP-IV VAR	4.9	4.6

Notes: Left: top row reports the true Smets and Wouters (2007) model parameters, and the remaining rows the mean estimates across 5000 Monte Carlo samples. All IV estimators use $h = 0, ..., H-1$ and the lag endogenous monetary policy instrument described in the text. SP-IV LP and LP-C denote implementations based on LPs without and with X_{t-1} (described in the text) as controls, respectively. SP-IV VAR denotes implementation with a VAR for X_t with four lags. Right: Tests for 2SLS/SE-IV use a Newey-West HAR variance matrix with Sun (2014) fixed-b critical values; inference procedures for SP-IV are described in Section 2 and Appendix A.

moments for identification.

The next two rows show the SP-IV estimators that condition on X_{t-1} using either LPs or a VAR. Both procedures produce mean estimates with the correct sign and values that are much closer to the truth. The reason for the smaller bias is the conditioning step, which helps eliminate the persistent influence of past markup shocks that leads to a violation of the lag exogeneity requirement. The lag-truncation bias present at longer horizons in the IRF coefficients underlying the SP-IV VAR estimates is relatively inconsequential. For $H = 20$, the SP-IV VAR estimates are, in fact, closer to the truth than the SP-IV LP-C estimates.

The right panel of Table 1 reports empirical rejection rates for nomi-

nal 5% tests that β equals the true value for the various SP-IV inference procedures described in Section 2 and Appendix A. Table 1 also reports analogous HAR procedures for the single equation specification with DL instruments.[11] When exogeneity fails, rejection rates will not match nominal levels. As Table 1 shows, every test based on exogeneity conditions that are violated (Wald 2SLS, AR/KLM SE-IV, and AR/KLM SP-IV LP) is indeed badly oversized. Conversely, for the SP-IV estimators that condition on X_{t-1}, (SP-IV LP-C and SP-IV VAR), the robust AR and KLM tests, defined in (20) and (21) respectively, exhibit empirical rejection rates very close to 5%, again demonstrating that the conditioning step adequately protects against the violation of lag exogeneity.[12] While the SP-IV estimators with controls have much smaller bias, some bias remains. The fact that robust inference procedures effectively control size indicates residual bias is related to the weakness of the instruments, even in a relatively large sample. This is consistent with the corresponding Wald test remaining somewhat oversized, especially when $H = 20$.

The results in Table 1 illustrate the advantage of SP-IV with controls relative to 2SLS in terms of weakening the exogeneity requirements. As explained in Section 1.3, adding X_{t-1} or X_{t-H} as controls to both stages of 2SLS does not similarly remove the lag endogeneity bias. Simulation results reported in the Online Appendix show that both these alternative 2SLS estimators perform very poorly under lag endogeneity. In our simulation setup, a version of 2SLS without controls but using a DL of $z_t^{\perp}$ – the residual in a regression of z_t on X_{t-1} – as instruments does successfully remove the endogeneity bias in large samples. The reason is that the Phillips curve residual u_t in the SW model does not depend on lags of the uncontaminated monetary policy shock. However, for the reasons given in Section 1.3, the same 2SLS estimator with a DL of $z_t^{\perp}$ as instruments does not perform as well as SP-IV with controls in simulations with small samples, which we discuss next.

[11] Both tests require HAR covariance estimates because of serial correlation in the Phillips curve residual u_t and the DL structure of the instrument set.

[12] As explained in Appendix A, the fact that IRF estimates at horizons beyond the VAR lag length are inconsistent does not affect the asymptotic validity of the AR or KLM tests for SP-IV VAR used in the simulations.

3.2. Small Sample Performance

Given the limited role of monetary policy shocks for inflation dynamics in the SW model, estimating the parameters of the Phillips curve using monetary policy shocks as instruments is especially challenging in small samples. The main goal of the next simulations is to show how the conditioning step in SP-IV may not only weaken exogeneity requirements but also substantially alleviate weak instrument problems. To level the playing field across estimators, we now assume that the econometrician has the true monetary policy shocks as instruments. This assumption is unrealistic but permits a comparison between the various estimators focused solely on instrument strength, as the exogeneity requirement is now satisfied for all IV estimators. We consider a sample of $T = 250$ quarters, a best-case scenario in most macro applications roughly corresponding to the postwar period, but also report results for $T = 500$ and $T = 5000$ to verify the asymptotic properties of the estimators and inference procedures. Note that our theoretical results are either asymptotic in nature or otherwise abstract from estimation error based on the inclusion of controls, so these simulations also present a laboratory to assess the trade-off of added estimation uncertainty in finite samples.

Bias. Table 2 reports the mean estimates of $\beta = [\gamma_b, \gamma_f, \lambda]'$ for various samples sizes. The first two rows report the true model parameters and OLS results. As expected, OLS is severely biased regardless of T due to endogeneity. The other rows show the results for 2SLS with DL instruments and the various SP-IV estimators with $H = 8$ or $H = 20$ quarters.

As the first row under $H = 8$ in Table 2 shows, 2SLS produces estimates that are closer on average to the true parameter values than OLS. Because the instruments are now lag exogenous, the 2SLS estimates converge to the truth as the sample size grows. However, despite the use of valid instruments, there remains considerable bias in realistic samples with $T = 250$. The Phillips Curve slope, λ, is estimated to be much flatter on average than in the model: 0.01 compared to 0.05. The backward and forward-looking inflation terms are also heavily misweighted, with γ_f too low on average and γ_b too high. The results are consistent with weak in-

Table 2: Mean parameter estimates

	$T = 250$			$T = 500$			$T = 5000$		
	γ_b	γ_f	λ	γ_b	γ_f	λ	γ_b	γ_f	λ
True Value	0.15	0.85	0.05	0.15	0.85	0.05	0.15	0.85	0.05
OLS	0.47	0.47	0.00	0.48	0.48	0.00	0.48	0.48	0.00
$H = 8$									
2SLS	0.27	0.51	0.01	0.24	0.61	0.00	0.17	0.83	0.04
SP-IV LP	0.26	0.51	0.01	0.24	0.60	0.00	0.17	0.83	0.04
SP-IV LP-C	0.29	0.64	0.05	0.25	0.74	0.04	0.16	0.84	0.05
SP-IV VAR	0.22	0.80	0.03	0.18	0.84	0.05	0.12	0.83	0.09
$H = 20$									
2SLS	0.39	0.53	0.00	0.36	0.61	0.00	0.23	0.80	0.01
SP-IV LP	0.38	0.53	0.00	0.35	0.61	0.00	0.23	0.80	0.01
SP-IV LP-C	0.41	0.55	0.01	0.37	0.64	0.01	0.23	0.81	0.02
SP-IV VAR	0.27	0.80	0.01	0.23	0.84	0.02	0.17	0.83	0.05

Notes: Top row reports the true parameter values in the Smets and Wouters (2007) model. The other rows report the mean estimates across 5000 Monte Carlo samples. All IV estimators are based on $h = 0, ..., H - 1$ and use true model shocks as instruments. SP-IV LP and LP-C denote implementations based on LPs without and with X_{t-1} as controls, respectively. SP-IV VAR denotes implementation with a VAR for X_t with four lags.

struments that bias 2SLS towards OLS. The next row shows that, without controls, the bias of SP-IV is almost identical to that of 2SLS for all T. This is again unsurprising as, in this case, both 2SLS and SP-IV exploit essentially the same identifying moments.

The next two rows under $H = 8$ illustrate the possible bias reductions when using the LP-C or VAR implementations of SP-IV, both of which condition on X_{t-1}. For the LP-C implementation, the estimates of λ average approximately the true value of 0.05 in samples with $T = 250$. The forward-looking coefficient in the Phillips Curve, γ_f, is also considerably closer to the truth, and the bias in the backward-looking coefficient, γ_b, is only marginally worse. The VAR implementation of SP-IV also delivers substantial bias improvements in all three coefficients. Relative to the LP-C implementation, the improvements are substantially larger for γ_b and γ_f, while the improvement for λ is somewhat smaller. Taken together,

the reductions in small sample bias by adopting SP-IV LP-C or SP-IV VAR are sizeable. These reductions are also economically meaningful, as the differences in mean parameter estimates have considerable implications for inflation dynamics and the inflation-output gap trade-off. As discussed in Section 1.3, the improvements relative to 2SLS arise because the conditioning step amplifies the signal provided by the monetary policy shock instrument, which is generally weak in the Smets and Wouters (2007) DGP. The Online Appendix presents additional results with multiple demand shocks as instruments ($N_z = 3$) that are qualitatively similar.

The extent of the improvements in the small sample performance of SP-IV relative to 2SLS depends on the choice of H. Including additional horizons can add useful identifying variation. On the other hand, the endogenous variables become harder to predict at longer horizons. The results in Table 2 for $H = 20$ show that the relative performance of the estimators is qualitatively the same as for $H = 8$. Quantitatively, however, the reductions in bias under the LP-C or VAR implementations of SP-IV are smaller than they are for $H = 8$. In general, as predicted in Section 1.3, the advantages of SP-IV over 2SLS diminish as the number of lags included as instruments in 2SLS – which is also the maximum forecast horizon in SP-IV – grows larger.

The Online Appendix shows that alternative versions of 2SLS that incorporate controls do not generate the same bias reductions as SP-IV with controls. Adding X_{t-1} as regressors in both stages greatly weakens identification and, on average, results in estimates of λ that have the wrong sign for all sample sizes. Regressing z_t on X_{t-1} and including a DL of the residual as instruments yields essentially the same results as 2SLS with a DL of z_t. Finally, including X_{t-H} as regressors in both stages leads to some improvement relative to 2SLS without controls, but the reductions in bias are meaningfully smaller than for SP-IV with controls.

Variance. Table 3 reports the standard deviations of the various IV estimators. Section 1.3 noted that SP-IV with controls is asymptotically more efficient than 2SLS when the error term u_t is an AR(1) process with $|\rho| > 0$. While the error term in our simulations is the ARMA(1,1) process

TABLE 3: STANDARD DEVIATION OF PARAMETER ESTIMATES

	$T = 250$			$T = 500$			$T = 5000$		
	γ_b	γ_f	λ	γ_b	γ_f	λ	γ_b	γ_f	λ
$H = 8$									
2SLS	0.27	0.33	0.20	0.24	0.30	0.21	0.13	0.08	0.09
SP-IV LP	0.27	0.35	0.22	0.25	0.30	0.22	0.13	0.08	0.09
SP-IV LP-C	0.28	0.27	0.25	0.26	0.20	0.23	0.12	0.06	0.08
SP-IV VAR	0.32	0.36	0.30	0.30	0.24	0.26	0.14	0.06	0.09
$H = 20$									
2SLS	0.11	0.12	0.05	0.10	0.11	0.06	0.07	0.05	0.03
SP-IV LP	0.12	0.13	0.06	0.11	0.11	0.06	0.07	0.05	0.03
SP-IV LP-C	0.09	0.11	0.06	0.09	0.09	0.05	0.08	0.05	0.04
SP-IV VAR	0.21	0.25	0.11	0.20	0.19	0.09	0.11	0.06	0.06

Notes: Standard deviations of the estimates across 5000 Monte Carlo samples from the Smets and Wouters (2007) model. All IV estimators are based on $h = 0, ..., H - 1$ and use true model shocks as instruments. SP-IV LP and LP-C denote implementations based on LPs without and with X_{t-1} as controls, respectively. SP-IV VAR denotes implementation with a VAR for X_t with four lags.

in (22), similar efficiency gains can arise. Table 3 indeed shows efficiency gains for $T = 5000$. For $H = 8$, the standard deviations of the SP-IV LP-C estimates are uniformly smaller than those of the 2SLS estimates. For the VAR implementation, the standard deviation is smaller for estimates of γ_f, and roughly similar to 2SLS for the other two parameters. Consistent with the theory, the relative efficiency of SP-IV diminishes for larger H, as can be seen for $H = 20$ and $T = 5000$ in the bottom panel.

Also consistent with the theory is that the conditioning step is essential to realize any efficiency gains: the SP-IV estimates that do not condition on X_{t-1}, in the second row of each panel, have similar or slightly larger variance than 2SLS. In smaller samples with $T = 250$ or 500, the LP-C implementation of SP-IV has some standard deviations smaller than 2SLS, and some larger. For $T = 250$ or 500, the standard deviations of SP-IV VAR, on the other hand, are systematically greater than 2SLS. The simulations also show that in small samples, there may be a trade-off to including controls due to estimation error in the forecast errors: for some parameters, there is increased standard deviation for SP-IV LP-C over

SP-IV LP for $H = 8, T = 250$ and $T = 500$.

At least for the DGP considered here, the LP-C implementation of SP-IV consistently generates lower bias than 2SLS, while it has similar or smaller variance. The VAR implementation yields further reductions in bias in our setting but generally also has slightly higher variance. That the VAR implementation has smaller bias but greater variance may be surprising given the bias-variance trade-off between VARs and LPs for the estimation of IRFs, see Li et al. (2021).[13] However, SP-IV does not estimate IRFs but relationships between IRFs. Biases and covariances across IRFs can have offsetting or reinforcing effects on the bias and variance of the SP-IV estimators. The balance of these effects, however, is application-specific.

Table 3 further shows that all standard deviations are decreasing in H, indicating that additional horizons reduce the variability of all IV estimators. Given our bias results, this implies a bias-variance trade-off when choosing the maximum horizon H for SP-IV: larger H provides smaller bias improvements relative to 2SLS with DL instruments but also generates less variable estimates. Simulations reported in the Online Appendix with three model shocks as instruments ($N_z = 3$) show a similar bias-variance trade-off for choosing N_z: using three instruments results in smaller bias improvements but also lower variances.

The Online Appendix also reports results for the feasible 2-step efficient GMM versions of SP-IV (or 'generalized' SP-IV), which are, in theory, asymptotically more efficient than our baseline estimators. Unfortunately, the feasible versions do not generally improve performance in practice, at least not for realistic sample sizes and our DGP.

A perhaps natural question is whether SP-IV offers advantages when multiple horizons are unnecessary to satisfy the order condition. In simulations available upon request, using $N_z = 6$ (all shocks except the cost-push shock), we found that the standard deviation of 2SLS estimates using only contemporaneous shocks was up to an order of magnitude higher than the SP-IV estimates using the same shocks and comparable to the SP-IV estimates using only $N_z = 1$. This is likely due to the hump-shaped nature of

[13] The Online Appendix shows that this trade-off is also present for the IRFs in our simulations.

IRFs in the Smets and Wouters (2007) model with contemporaneous responses that are fairly muted, resulting in very weak identification. SP-IV is therefore appealing even when many shocks are available and $N_z \geq K$.

Inference. Given that monetary policy shocks are weak instruments in the Smets and Wouters (2007) DGP and realistic sample sizes, a key question is how severe size distortions are using standard Wald inference and how well the weak instrument robust procedures control size in practice. It is well known that robust procedures may still perform poorly when the number of instruments is large (Bekker 1994). Barnichon and Mesters (2020), for example, report severe size distortions for AR inference with long lag sequences of instrumenting shocks. Since SP-IV uses HN_z moments, it potentially faces the same theoretical "many-moments" problem as 2SLS with HN_z instruments (Han and Phillips 2006; Newey and Windmeijer 2009).

Table 4 reports empirical rejection rates for nominal 5% tests of the true values of the full parameter vector, $\beta = [\gamma_b, \gamma_f, \lambda]'$ for sample sizes of $T = 250, 500$, and 5000. To better assess distortions due to many moments, Table 4 also reports results for $N_z = 3$ and $H = 20$; see the Online Appendix for details. Note that size distortions related to weak instruments will generally decrease with T since the first-stage relationships remain fixed, and identification strength improves with T.

As Table 4 shows, Wald tests for 2SLS exhibit meaningful size distortions for $H = 8$, with empirical rejection rates substantially above the nominal 5%. The size distortions increase meaningfully as H and/or N_z become larger. These distortions are not surprising given the weakness of the instruments and demonstrate the need for robust inference procedures. For $H = 8$ and $N_z = 1$, the conventional AR test for SE-IV is relatively well-sized in small samples. Indicative of many-moment problems, the AR SE-IV test becomes noticeably oversized in small samples when $H = 20$, and even more so when in addition $N_z = 3$. The KLM test for SE-IV controls size better but can be conservative, which is potentially due to the use of only approximate fixed-b critical values for HAR inference.

Just like for 2SLS, the SP-IV Wald tests show size distortions that

TABLE 4: EMPIRICAL SIZE OF NOMINAL 5% TESTS

	$H = 8$, $N_z = 1$			$H = 20$, $N_z = 1$			$H = 20$, $N_z = 3$		
$T =$	250	500	5000	250	500	5000	250	500	5000
Wald 2SLS	9.1	7.5	10.9	67.2	60.8	45.4	100.0	99.9	94.3
Wald SP-IV LP	16.4	12.9	14.3	71.2	67.0	47.7	100.0	99.9	93.8
Wald SP-IV LP-C	14.8	12.5	9.3	74.6	67.0	34.9	100.0	99.8	83.0
Wald SP-IV VAR	7.9	6.5	5.5	33.7	27.0	13.3	86.7	76.7	54.1
AR SE-IV	7.6	6.8	4.6	14.0	10.5	5.2	60.0	36.3	6.4
AR SP-IV LP	6.3	5.7	5.0	10.4	6.9	5.8	14.3	8.0	5.0
AR SP-IV LP-C	6.4	5.7	4.6	11.3	7.1	5.8	16.9	9.2	5.1
AR SP-IV VAR	4.2	4.9	4.9	5.8	5.6	4.5	6.5	5.2	4.6
KLM SE-IV	3.9	3.7	3.8	4.7	5.1	4.9	0.0	7.2	5.0
KLM SP-IV LP	6.0	5.2	5.0	7.5	6.5	4.9	7.6	6.5	5.3
KLM SP-IV LP-C	7.4	5.4	5.2	11.2	7.4	5.5	11.4	7.6	6.1
KLM SP-IV VAR	5.6	5.3	4.9	8.5	6.8	4.6	11.7	8.5	5.5

Notes: Empirical rejection rates of nominal 5% tests of the true values of $\beta = [\gamma_b, \gamma_f, \lambda]'$ in 5000 Monte Carlo samples from the Smets and Wouters (2007) model. All IV estimators are based on $h = 0, ..., H-1$ and use true model shocks as instruments. Tests for 2SLS/SE-IV use a Newey-West HAR variance matrix with Sun (2014) fixed-b critical values; inference procedures for SP-IV are described in Section 2 and Appendix A.

become very large as H and N_z increase. The SP-IV AR tests are overall well-sized. Both the LP and LP-C implementations over-reject in small samples when $H = 20$, but somewhat less so than AR SE-IV. The SP-IV KLM tests are also generally well-sized. Just like the AR tests, the KLM tests exhibit some over-rejection in small samples when $H = 20$ and/or $N_z = 3$. Overall, however, the size distortions of the robust SP-IV tests with large HN_z appear milder than those for the SE-IV versions.[14] Because of many-moment problems, we nevertheless recommend avoiding very large HN_z also when using SP-IV. Even with a single instrument, the rejection rates of the robust tests likely approach many applied researchers' tolerance for size distortions for $T = 250$ and $H = 20$. While based on just a single DGP, $H = 20$ therefore appears as an upper bound for $T = 250$ in this setting, with the optimal number of horizons likely

[14] One likely reason that the robust SP-IV procedures control size distortions better is that SE-IV requires HAR covariance estimates, see footnote 11. As a result, the number of parameters to be estimated increases much more quickly with HN_z than for the SP-IV covariances.

considerably lower.

In practice, there is fortunately no need to use all horizons for identification. Researchers can, for example, select impulse response horizons at lower frequencies than that of the time series (e.g., quarterly horizons in monthly data, annual horizons in quarterly data, etc.), especially since adjacent horizons do not necessarily contain much independent identifying information for typical shapes of IRFs. Further refinements are also possible to address any remaining many instrument problems, see for example Mikusheva (2021) for suggestions. In the context of 2SLS with DL instruments, Barnichon and Mesters (2020) propose quadratic approximations to the IRFs to avoid many instrument problems, and similar approximations are possible with SP-IV.[15] Other test statistics could possibly be adapted to SP-IV and offer improvements over the AR and KLM tests, for example, those based on Moreira (2003) or Andrews (2016). Given the relatively good performance of our robust test statistics in the simulations, we leave such extensions for future work.

4. Application to the Phillips Curve with U.S. Data

In this section, we use SP-IV to estimate the parameters of the Phillips curve in (2) using U.S. data and compare the results with 2SLS with DL instruments. We consider the following specification for monthly inflation,

$$\pi_t^{1m} = (1-\gamma_f)\pi_{t-1}^{1y} + \gamma_f \pi_{t+12}^{1y} + \lambda U_t + u_t \ , \tag{23}$$

where π_t^{1m} is the annualized monthly percent change in the Core CPI, π_t^{1y} is the percent change in the Core CPI over the preceding year in month t, and U_t is the headline unemployment rate in month t. The specification and variable definitions are similar to Barnichon and Mesters (2020), but we use monthly data from Jan 1978 to Feb 2020 (506 observations) instead of quarterly data. As is common in the literature, e.g., Mavroeidis et al.

[15] In simulations available on request, we consider Barnichon and Mesters's (2020) 2SLS estimator with Almon shrinkage. The performance is poor, with bias highly variable over H, T, and parameters, and standard deviations one to two orders of magnitude larger than those of the other estimators. No inference procedure controls size well across all specifications. In contrast, we find that SP-IV LP-C and VAR perform very well when tested on Barnichon and Mesters's (2020) DGP.

(2014), (23) restricts the coefficients on lagged and future inflation to sum to unity, $\gamma_b + \gamma_f = 1$, which imposes that there is no long-run trade-off between unemployment and inflation. The restriction is implemented by rewriting (23) as $\pi_t^{1m} - \pi_{t-1}^{1y} = \gamma_f(\pi_{t+12}^{1y} - \pi_{t-1}^{1y}) + \lambda U_t + u_t$. We consider a maximum forecast horizon of 3 years (36 months). To make efficient use of the identifying information in the IRF dynamics and mitigate many-moment problems, we only use the coefficients in the first month of each of the first 12 quarters of the response horizons – that is, $h = 0, 3, 6, \ldots, 33$. We consider identification with a single economic shock, such that for both SP-IV and 2SLS there are 12 identifying moments. We use the VAR implementation of SP-IV, using a VAR with six lags in the following standard monthly macro variables as controls: the annualized monthly percent change in the core CPI, the unemployment rate, the 12-month change in log industrial production, the 12-month percent change in the PPI for all commodities, the 3-month Treasury rate, and the 10-year Treasury rate.

As the instrument, we use a monthly version of the Angeletos et al. (2020) Main Business Cycle (MBC) Shock, identified within our VAR by maximizing the contribution to cyclical unemployment fluctuations in the frequency domain. Angeletos et al. (2020) find that the resulting shock is interchangeable with shocks identified by maximizing the cyclical variance contribution to other major macro aggregates, such as GDP, consumption, investment, or hours worked. This interchangeability suggests a single main driver of business cycles with a common propagation mechanism. Empirically, this propagation mechanism best fits the notion of an aggregate demand shock, in which case the MBC shock is a valid instrument for estimating the Phillips curve.

While one can certainly question the validity of the MBC shock as an instrument, we chose it for two main reasons. The first reason is that it serves as a good illustration of how SP-IV can be useful when interpreting empirical IRFs. Observing the disconnect between the unemployment and inflation impulse responses to the MBC shock, a key conclusion in Angeletos et al. (2020) is that the Phillips curve must be nearly completely flat. Rather than relying on informal visual inspections of IRFs, SP-IV allows a formal econometric investigation of the Phillips curve relation-

ship embedded in the VAR-based IRFs. The second reason is that the MBC shock is likely the strongest available instrument for identifying the Phillips curve. By construction, the MBC shock is highly predictive of unemployment fluctuations over business cycle horizons. We find that the MBC shock also has some strength for inflation at horizons up to three years. The first two columns in the first row of Table 5 show test statistics and critical values of the weak instruments test for SP-IV, along with those for 2SLS (without controls) based on the HAR first-stage test of Lewis and Mertens (2022). For illustrative purposes, the other columns in Table 5 report results for each endogenous regressor separately. The test statistic is 6.3 for SP-IV, with a critical value of 22.0 for the null of relative bias of at most 10% at the 5% significance level (the threshold suggested by Stock and Yogo (2005)). The test statistics are 19.8 and 6.4, respectively, for unemployment and inflation separately (critical values of 21.7 and 18.6). The test statistics for 2SLS are all much lower relative to similar critical values of around 20, which illustrates how including the additional predictors in SP-IV amplifies the signal of the instrument relative to 2SLS. Despite this amplification, the MBC shock is still judged to be weak at conventional tolerance levels according to the SP-IV first-stage test. The MBC shock is, nevertheless, by far the strongest instrument across all candidate shocks that we explored.

In principle, many other shock measures could be used to identify the parameters of (23), including monetary policy shocks as in Barnichon and Mesters (2020). Table 5 reports the first stage test results for various popular monetary policy shock measures. The main takeaway is that, at least in the sample that we consider, each of the monetary policy shocks is far too weak as an instrument to be useful for identifying the Phillips curve in practice. A few have some strength for inflation separately in the first stage of 2SLS, but none do for both endogenous regressors jointly, which is what matters for identification. Moreover, any hint of instrument strength disappears entirely after including lagged macroeconomic variables as controls in SP-IV. That none of the 2SLS/SP-IV first-stage tests with monetary policy shocks comes close to rejecting the null of weak instruments is not surprising, as it reflects the broadly held view

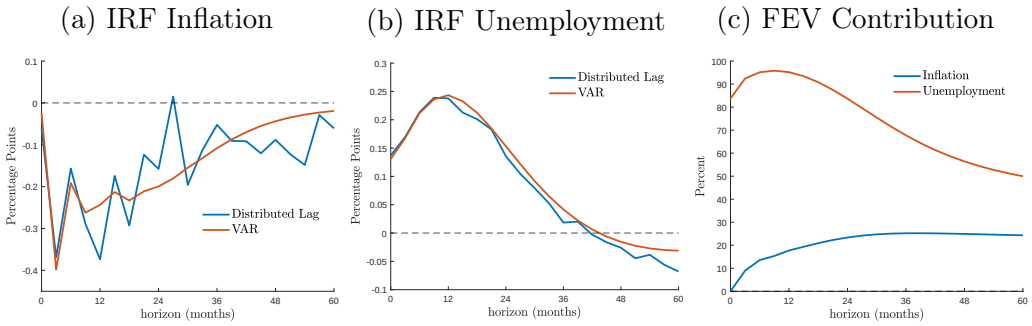
TABLE 5: FIRST-STAGE TEST RESULTS

	π, U jointly				U separately				π separately			
	2SLS		SP-IV		2SLS		SP-IV		2SLS		SP-IV	
	g	cv	g	cv	g	cv	g	cv	g	cv	g	cv
MBC	2.3	21.9	6.3	22.0	4.2	20.4	19.8	21.7	2.3	19.5	6.4	18.6
Monetary Policy Shock Measures:												
RR	0.3	20.2	0.4	21.2	0.3	17.7	0.5	20.0	1.9	17.8	0.9	18.6
GK	3.5	21.0	0.1	22.3	3.5	20.3	0.4	22.1	7.4	18.8	0.1	17.3
MAR	0.1	19.7	0.0	21.0	0.2	17.8	0.1	20.9	0.8	18.6	0.0	17.3
JK	0.6	19.9	0.0	22.3	1.5	18.9	0.2	22.1	2.0	18.6	0.1	17.3
SWA	1.0	19.7	0.0	22.3	1.0	19.4	0.1	22.1	5.7	17.2	0.0	17.3
BC	1.3	21.1	0.0	22.3	2.1	19.7	0.1	22.1	5.4	18.9	0.0	17.3
SN	1.0	20.2	0.0	22.5	1.1	19.4	0.1	22.1	7.1	17.0	0.0	17.6

Notes: The table reports test results for the null hypothesis of weak instrument bias less than or equal to 10% of the worst-case benchmark. g is the test statistic, cv is the 5% critical value. U and π are the endogenous regressors in the restricted equation, i.e. U_t and $\pi^{1y}_{t+12} - \pi^{1y}_{t-3}$. For 2SLS, results are for the HAR test of Lewis and Mertens (2022). For SP-IV, the test is described in the Online Appendix. MBC is the main business cycle shock of Angeletos et al. (2020). The monetary policy shock measures are Romer and Romer (2004) (RR), Gertler and Karadi (2015) (GK), Miranda-Agrippino and Ricco (2021) (MAR), Jarociński and Karadi (2020) (JK), Swanson (2021) (SWA), Barakchian and Crowe (2013) (BC), Nakamura and Steinsson (2018) (NS). RR, GK, and BC are updated versions from Ramey (2016); the other series are from the original sources.

that monetary disturbances are relatively unimportant as drivers of inflation and economic activity. We also considered several other plausible demand shock measures, such as the credit spread shock of Gilchrist and Zakrajšek (2012) and the Bloom (2009) uncertainty shock, but none are nearly as strong as the MBC shock in the post-1978 sample. Using multiple shocks could improve identification strength but creates potential many-instrument problems. To the extent that the MBC shock indeed

Figure 1: Impact of the MBC Shock on Inflation and Unemployment



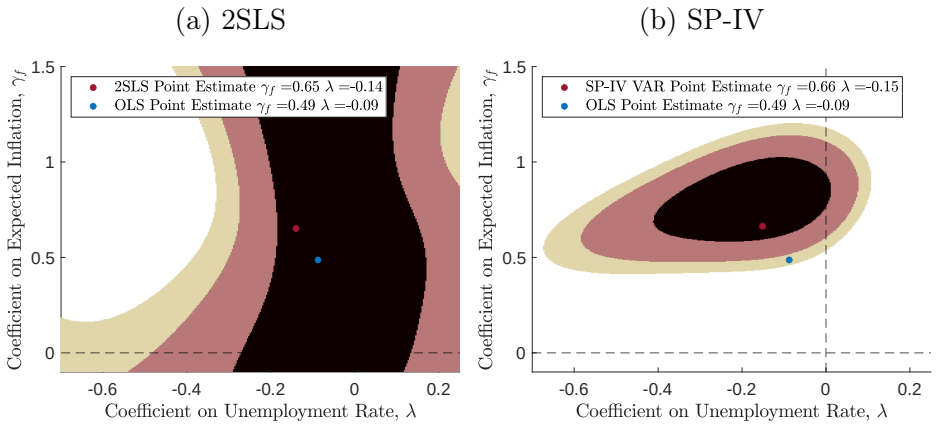
Notes: Inflation is the annualized Core CPI inflation rate from a quarter ago (π_t^{1q}). IRFs in red in (a) and (b) are from a VAR with six lags using the annualized monthly core CPI inflation, the unemployment rate, the 12-month change in log industrial production, the 12-month percent change in the PPI for all commodities, the 3-month Treasury rate, and the 10-year Treasury rate. The estimation sample is from Jan 1978 to Feb 2020. Blue lines in (a) and (b) show results from the DL regressions in the first stage of 2SLS. Panel (c) shows the FEV contributions of the MBC shock in the VAR.

collects a range of demand disturbances that satisfy the exogeneity requirements, it is by far the most informative available instrument for the identification of the Phillips curve.

Our monthly version of the MBC shock produces IRFs that are consistent with those in Angeletos et al. (2020). The red lines in Figures 1a-1b plot VAR-based IRFs of π_t^{1m} and U_t to a one-standard-deviation MBC shock. The figures show the first twelve IRF coefficients that are used in the estimation and also show the next eight quarters to visualize the full dynamics. As in Angeletos et al. (2020), the MBC shock looks like an aggregate demand shock, driving unemployment higher and inflation lower. At the same time, the MBC shock explains a relatively small fraction of the forecast error variance (FEV) of inflation, nearly zero on impact and only 20% after two years, see Figure 1c. This finding illustrates the apparent "disconnect" between inflation and the shock that explains most of the cyclical variation in unemployment, see also Del Negro et al. (2020). As previously explained, both 2SLS and SP-IV estimates can be expressed as the coefficients in regressions of IRFs. In SP-IV, these IRFs are the red lines in Figures 1a-1b.[16] The 2SLS estimator instead uses the IRFs

[16]The IRF of $\pi_{t+12}^{1y} - \pi_{t-1}^{1y}$ is straightforward to construct from the IRF of π_t^{1m}.

Figure 2: 2SLS and SP-IV Confidence Sets for Estimates of Phillips Curve Parameters



Notes: Figures show point estimates and 68%, 90% and 95% confidence sets based on the KLM statistic for SP-IV VAR described in Appendix A. Confidence sets for 2SLS are based on the KLM test and a Newey-West HAR covariance matrix with Sun (2014) fixed-b critical values.

obtained from DL regressions of π_t^{1m} (and π_{t-1}^{1y} and π_{t+12}^{1y}) and U_t on the current and lagged values of the MBC shock. For illustration, these IRFs are shown in blue in Figures 1a-1b.

Figure 2 displays the estimates of γ_f and λ, together with 68%, 90% and 95% confidence sets. Since neither of the first-stage tests in Table 5 rejects the null of weak instruments, the confidence sets are both based on the KLM statistic. The point estimates of γ_f, the weight on future inflation, are 0.65 for 2SLS and 0.66 for SP-IV. The slope estimates are also close, $\lambda = -0.15$ in SP-IV versus $\lambda = -0.14$ in 2SLS, and have the expected negative sign since unemployment is the gap measure. The similarity in point estimates is not too surprising, given that we use the VAR-identified MBC shock to construct the instruments for 2SLS. The inference results, on the other hand, are much less similar. The confidence sets based on the 2SLS's SE-IV moments do not reject any plausible values of γ^f, nor do they rule out a wide range of possible values of λ. Comparatively, inference for SP-IV is much sharper for the weights on inflation, with the confidence set ruling out values of γ_f that are meaningfully below 0.5 or above 1. At the same time, the SP-IV sets also do not rule

out a wide range of possible Phillips curve slopes, with values of λ ranging from close to -0.6 to somewhat greater than zero within the 90% set. We attribute the relatively more informative SP-IV confidence sets to the greater effective strength of the instruments, as discussed in Section 1.3.

As to the inflation-activity disconnect, our robust inference results act as a warning against drawing strong conclusions from informal comparisons of IRF point estimates. When judging relationships across IRFs, it is important to take into account that these estimates are inevitably uncertain. SP-IV estimates the posited relationships between IRFs from VARs or LPs formally and allows inference that is robust to the distortions caused by sampling error in the IRF estimates. The confidence sets in Figure 2b, for example, are consistent with weak but also relatively strong cyclical connections between inflation and unemployment. The business cycle anatomy of Angeletos et al. (2020), therefore, does not provide clear-cut evidence that inflation and activity are largely disconnected.

5. Concluding Remarks and Future Research

While we focused mainly on identifying the parameters of the inflation Phillips curve, SP-IV can help identify a wide variety of structural relationships in macroeconomics, such as Euler equations for consumption or investment, the wage Phillips curve, monetary or fiscal policy rules, and aggregate production functions.[17] SP-IV can be used more broadly to conduct inference on ratios (or other relationships) of impulse response coefficients, such as Okun coefficients, sacrifice ratios, multipliers, etc., conditional on economic shocks. Our methodology could be extended to panel data settings and should be more generally useful in applications that commonly rely on lagged variables as instruments, such as the estimation of production functions in industrial organization. SP-IV could also be used in cross-sectional applications. If $h = 0, ..., H-1$ indexes cross-sectional groups rather than time horizons, then SP-IV amounts to instrumental variables in the cross-section with heterogeneity in the first

[17] Fieldhouse and Mertens (2023), for example, use SP-IV to estimate the aggregate output elasticity to public R&D capital using IRFs to shocks to federal appropriations for R&D funding.

stage coefficients. Future work can also develop Bayesian implementations or methods to select the horizons/groups used for identification optimally.

References

Anderson, T. W., & 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, I. (2016). Conditional Linear Combination Tests for Weakly Identified Models. *Econometrica*, *84*(6), 2155–2182.

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

Angeletos, G.-M., Collard, F., & Dellas, H. (2020). Business-Cycle Anatomy. *American Economic Review*, *110*(10), 3030–70.

Barakchian, S. M., & Crowe, C. (2013). Monetary policy matters: Evidence from new shocks data. *Journal of Monetary Economics*, *60*(8), 950–966.

Barnichon, R., & Mesters, G. (2020). Identifying Modern Macro Equations with Old Shocks. *The Quarterly Journal of Economics*, *135*(4), 2255–2298.

Bauer, M. D., & Swanson, E. T. (2022). *A Reassessment of Monetary Policy Surprises and High-Frequency Identification* (Working Paper No. 29939). National Bureau of Economic Research.

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

Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica*, *77*(3).

Cieslak, A. (2018). Short-Rate Expectations and Unexpected Returns in Treasury Bonds. *The Review of Financial Studies*, *31*(9), 3265–3306.

Coibion, O. (2012). Are the Effects of Monetary Policy Shocks Big or Small? *American Economic Journal: Macroeconomics*, *4*(2), 1–32.

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

Del Negro, M., Lenza, M., Primiceri, G. E., & Tambalotti, A. (2020). What's up with the Phillips Curve? *Brookings Papers on Economic Activity, Spring*.

Fieldhouse, A. J., & Mertens, K. (2023). *The Returns to Government RD: Evidence from U.S. Appropriations Shocks* (Working Paper No. 2305). Federal Reserve Bank of Dallas.

Galí, J., & Gambetti, L. (2020). Has the U.S. Wage Phillips Curve Flattened? A Semi-Structural Exploration. In G. Castex, J. Galí, & D. Saravia (Eds.). Central Bank of Chile.

Galí, J., & Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics*, *44*(2), 195–222.

Gertler, M., & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, *7*(1), 44–76.

Gilchrist, S., & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review*, *102*(4), 1692–1720.

Han, C., & Phillips, P. C. B. (2006). GMM with Many Moment Conditions. *Econometrica*, *74*(1), 147–192.

Jarociński, M., & Karadi, P. (2020). Deconstructing Monetary Policy Surprises—The Role of Information Shocks. *American Economic Journal: Macroeconomics*, *12*(2), 1–43.

Jordà, (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, *95*(1), 161–182.

Jordà, & Kozicki, S. (2011). Estimation and Inference by the Method of Projection Minimum Distance: An Application to the New Keynesian Hybrid Phillips Curve. *International Economic Review*, *52*(2), 461–487.

Kilian, L., & Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.

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.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics*, *47*(3), 523–544.

Lewis, D. J., & Mertens, K. (2022). *A Robust Test for Weak Instruments with Multiple Endogenous Regressors* (Staff Reports No. 1020). Federal Reserve Bank of New York.

Li, D., Plagborg-Moller, M., & Wolf, C. K. (2021). *Local Projections vs. VARs: Lessons From Thousands of DGPs* (Papers No. 2104.00655). arXiv.org.

Lloyd, S., & Manuel, E. (2023). *Controls, Not Shocks: Estimating Dynamic Causal Effects in the Face of Confounding Factors* (tech. rep.).

Mavroeidis, S., Plagborg-Møller, M., & Stock, J. H. (2014). Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve. *Journal of Economic Literature*, *52*(1), 124–88.

McLeay, M., & Tenreyro, S. (2019). Optimal Inflation and the Identification of the Phillips Curve. *NBER Macroeconomics Annual*, *34*.

Mertens, K., & Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. *American Economic Review*, *103*(4), 1212–47.

Mikusheva, A. (2021). *Many Weak Instruments in Time Series Econometrics* (World Congress of Econometric Society). MIT.

Miranda-Agrippino, S., & Ricco, G. (2021). The Transmission of Monetary Policy Shocks. *American Economic Journal: Macroeconomics*, *13*(3), 74–107.

Montiel-Olea, J. L., & Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, *31*(3), 358–369.

Montiel Olea, J. L., & Plagborg-Møller, M. (2021). Local Projection Inference is Simpler and More Robust Than You Think. *Econometrica*, *89*(4), 1789–1823.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, *71*(4), 1027–1048.

Nakamura, E., & Steinsson, J. (2018). High-Frequency Identification of Monetary Non-Neutrality: The Information Effect. *The Quarterly Journal of Economics*, *133*(3), 1283–1330.

Newey, W. K., & Windmeijer, F. (2009). Generalized Method of Moments with Many Weak Moment Conditions. *Econometrica*, *77*(3), 687–719.

Plagborg-Møller, M., & Wolf, C. K. (2021). Local Projections and VARs Estimate the Same Impulse Responses. *Econometrica*, *89*(2), 955–980.

Plagborg-Møller, M., & Wolf, C. K. (2022). Instrumental variable identification of dynamic variance decompositions. *Journal of Political Economy*, *130*(8), 2164–2202.

Ramey, V. (2016). Macroeconomic Shocks and Their Propagation. In J. B. Taylor & H. Uhlig (Eds.). Elsevier.

Romer, C. D., & Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, *94*(4), 1055–1084.

Rothenberg, T. J., & Leenders, C. T. (1964). Efficient Estimation of Simultaneous Equation Systems. *Econometrica*, *32*(1/2), 57–76.

Smets, F., & Wouters, R. (2007). Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. *American Economic Review*, *97*(3), 586–606.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. In D. W. Andrews (Ed.), *Identification and Inference for Econometric Models* (pp. 80–108). Cambridge University Press.

Stock, J. H., & Watson, M. W. (2012). Disentangling the Channels of the 2007-2009 Recession. *Brookings Papers on Economic Activity, Spring 2012*, 81–135.

Stock, J. H., & Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, *128*(610), 917–948.

Stock, J. H., & Wright, J. H. (2000). GMM with Weak Identification. *Econometrica*, *68*(5), 1055–1096.

Sun, Y. (2014). Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics*, *178*(P3), 659–677.

Swanson, E. T. (2021). Measuring the Effects of Federal Reserve Forward Guidance and Asset Purchases on Financial Markets. *Journal of Monetary Economics*, *118*, 32–53.

Appendix

A. Practical Implementation of SP-IV with LPs or VARs

Let y_H denote the $H \times T$ matrix of leads of the outcome variable, i.e. with y_{t+h} in the $h+1$-th row and t-th column. Let Y_H be the $HK \times T$ matrix vertically stacking the $H \times T$ matrices Y_H^k for $k = 1, \ldots, K$, each of which has Y_{t+h}^k in the $h+1$-th row and t-th column, and Y_t^k the k-th variable in the vector Y_t. Let X_t be the period t observation of an $N_x \times 1$ collection of predetermined control variables (including a constant). X_t can include not only current values, but also lags of y_t, Y_t, z_t, or any other time series. Since the focus is computing forecast errors, we now make estimated forecast errors explicit using $\hat{y}_H^\perp$, etc.

Local Projections Define the $N_x \times T$ matrix X with controls X_{t-1} in the t-th column, the projection matrix $P_X = X'(XX')^{-1}X$, and residualizing matrix $M_X = I_T - P_X$. Using a direct forecasting approach, the forecast errors after projection on X_{t-1} are given by

$$\text{(A.1)} \qquad \hat{y}_H^\perp = y_H M_X \quad , \quad \hat{Y}_H^\perp = Y_H M_X \quad , \quad \hat{Z}^\perp = Z M_X \ ,$$

which can be used in (9) to obtain the SP-IV estimator $\hat{\beta}$. By the Frisch-Waugh-Lovell Theorem, this direct forecasting approach is equivalent to estimating Jordà (2005) local projections of y_{t+h} and Y_{t+h} on z_t and X_{t-1} for $h = 0, \ldots, H-1$, using the estimated coefficients on z_t to construct the rows of $\hat{\mathbf{\Theta}}_y$ and $\hat{\mathbf{\Theta}}_Y$ and subsequently constructing the SP-IV estimator using the alternative expression for $\hat{\beta}$ in (13). When $\hat{Z}^\perp$ are measures of economic shocks, the LP estimates are IRF coefficients representing the dynamic causal effects of the shocks. Some studies estimate IRFs by local projections of an endogenous outcome variable at $t+h$ on an endogenous explanatory variable Y_t^k and controls X_{t-1} using z_t as instruments, a procedure often referred to as "LP-IV". Such IRFs can be used for identification in the SP-IV estimator exactly as described above, i.e. using the reduced form projections of the outcome variables on z_t and X_{t-1}.

Vector Autoregressions Suppose that y_t, Y_t, and z_t, are – possibly together with other variables – all contained in X_t and that X_t evolves according to a VAR,

$$X_t = AX_{t-1} + e_t \ . \tag{A.2}$$

The representation in terms of a VAR of order one is without loss of generality, as any VAR of order p can be rewritten as a VAR of order one (in "companion form"). As before, let X denote the $N_x \times T$ matrix with X_{t-1} in the t-th column, and let X^f denote the $N_x \times T$ matrix with X_t in the t-th column. The standard estimator of A is $\hat{A} = X^f X'(XX')^{-1}$, leading to the h-step ahead forecast errors

$$\hat{X}_t^{\perp}(h) = \sum_{j=0}^{h} \hat{A}^{h-j}\hat{e}_{t+j} \quad , \quad \hat{e}_t = X_t - \hat{A}X_{t-1} \ . \tag{A.3}$$

The appropriate selection of elements in $\hat{X}_t^{\perp}(h)$ leads to $\hat{y}_H^{\perp}$, $\hat{Y}_H^{\perp}$ and $\hat{Z}^{\perp}$, which can be used to obtain the SP-IV estimator $\hat{\beta}$ in (9). Note that if Y_t contains lagged or lead variables, as in the HNKPC, only the contemporaneous value must be included in the VAR, and the estimated IRFs themselves can be time-shifted forwards or backwards when constructing $\hat{\mathbf{\Theta}}_Y$. "Structural" VARs are VARs in which researchers make assumptions to identify columns of D in $e_t = D\epsilon_t$, allowing the estimation of IRFs that are interpretable as dynamic causal effects of the associated economic shocks in ϵ_t. If $\hat{\epsilon}_t^{1:N_z}$ are the N_z identified shocks in the structural VAR, it is possible to use $\hat{z}_t^{\perp} = \hat{\epsilon}_t^{1:N_z}$ to form $\hat{Z}^{\perp}$ and use these shock estimates for identification in the SP-IV estimator. This procedure also nests identification with "external instruments", which can be directly included in the VAR and combined with zero restrictions in D as proposed by Plagborg-Møller and Wolf (2021), or used indirectly as instruments to identify columns in D as in the "proxy SVAR" or "SVAR-IV" approach (Mertens and Ravn 2013; Stock and Watson 2012; Stock and Watson 2018). Note that (11), or equivalently (12), are consistent estimators of the IRFs associated with $\hat{\epsilon}_t^{1:N_z}$ if the VAR restrictions imposed hold in the DGP. In finite samples, however, these IRF estimates will not be numeri-

cally identical to those obtained from $\hat{\Theta}_{X,h}^{VAR} = \hat{A}^h D^{1:Nz}$, $h = 0, \ldots, H-1$, where $D^{1:Nz}$ denotes the first N_z columns of D. The reason is that the restrictions implied by the VAR dynamics are imposed on the reduced form forecast errors, but (11) or (12) do not impose the same VAR dynamics on the IRFs.

Our preferred implementation of SP-IV with structural VARs is instead to select the elements corresponding to y_t and Y_t in $\hat{\Theta}_{X,h}^{VAR}$ to form $\hat{\mathbf{\Theta}}_y$ and $\hat{\mathbf{\Theta}}_Y$, and then obtain the SP-IV estimator from the regression of impulse responses as in (13). This alternative implementation imposes the VAR dynamics on both the reduced form forecast errors and the impulse responses. In general, imposing the VAR dynamics is easily done in all formulas above by replacing $\hat{y}_H^\perp P_{\hat{Z}^\perp} \hat{Y}_H^\perp$ by $\hat{\Theta}_y^{VAR}(\hat{\Theta}_Y^{VAR})'$ and $\hat{Y}_H^\perp P_{\hat{Z}^\perp} \hat{Y}_H^\perp$ by $\hat{\Theta}_Y^{VAR}(\hat{\Theta}_Y^{VAR})'$, where $\hat{\Theta}_Y^{VAR}$ is the $HK \times N_z$ matrix stacking the K blocks of the VAR IRF coefficients of Y_t, and $\hat{\Theta}_y^{VAR}$ contains the $H \times N_z$ VAR IRF coefficients of y_t. When comfortable imposing VAR dynamics, it makes sense to impose these restrictions consistently, and we therefore recommend this second implementation for SP-IV with VARs.

To impose the VAR dynamics in the Generalized SP-IV formula (B.1), replace $\hat{y}_H^\perp P_{\hat{Z}^\perp}$ by $\hat{\Theta}_y^{VAR}(ZM_XZ'/T)^{-\frac{1}{2}}ZM_X$. To impose the VAR dynamics in the KLM statistic in (21), replace $\hat{Y}_H^\perp P_{\hat{Z}^\perp}$ by $\hat{\Theta}_Y^{VAR}(ZM_XZ'/T)^{-\frac{1}{2}}ZM_X$, replace $\hat{u}_H^\perp(b)P_{\hat{Z}^\perp}$ by $\left(\hat{\Theta}_y^{VAR} - (b' \otimes I_H)\hat{\Theta}_Y^{VAR}\right)(ZM_XZ'/T)^{-\frac{1}{2}}ZM_X$, and $\hat{u}_H^\perp(b)M_{\hat{Z}^\perp}$ by $\hat{u}_H^\perp(b) - \left(\hat{\Theta}_y^{VAR} - (b' \otimes I_H)\hat{\Theta}_Y^{VAR}\right)(ZM_XZ'/T)^{-\frac{1}{2}}ZM_X$. When imposing the VAR restrictions in the AR test for SP-IV, some additional care is warranted. First, when replacing $\hat{u}_H^\perp(b)P_{\hat{Z}^\perp}$ in (20), it is important that the VAR IRFs are consistent for all horizons. This requires that the VAR restrictions hold in the DGP for inference based on the AR statistic to be a correctly sized test of the null hypothesis. Second, the normalizing variance in (20) must also take the VAR restrictions into account, which can be done in practice by computing the variance using the Delta Method.

As explained in the main text, the SW model used in the simulations of Section 3 does not permit a finite-order VAR representation in X_t. As a result, the IRFs based on a VAR with four lags are not (asymptotically) unbiased for horizons exceeding the lag length, see Plagborg-Møller

TABLE A.1: SPIV-VAR: EMPIRICAL SIZE OF NOMINAL 5% TESTS

	$H = 8$, $N_z = 1$			$H = 8$, $N_z = 3$		
$T =$	250	500	5000	250	500	5000
AR SP-IV VAR, FE only	4.2	4.9	4.9	3.9	5.1	4.8
AR SP-IV VAR, IRFs and FE	4.7	3.1	7.6	3.4	1.9	91.7
AR SP-IV VAR, IRFs and FE, 8 lags	25.2	11.7	5.7	49.2	19.2	6.1
KLM SP-IV VAR, FE only	5.5	5.3	5.1	5.9	5.6	4.7
KLM SP-IV VAR, IRFs and FE	5.6	5.3	4.9	6.9	6.6	4.9
KLM SP-IV VAR, IRFs and FE, 8 lags	5.0	5.1	5.0	6.7	6.0	4.7

Notes: Empirical rejection rates of nominal 5% tests of the true values of $\beta = [\gamma_b, \gamma_f, \lambda]'$ in 5000 Monte Carlo samples from the Smets and Wouters (2007) model. All results are for SP-IV based on a VAR in X_t using true model shocks as the instrument. 'FE only' uses the VAR only to obtain forecast errors, but does not impose the VAR restrictions on the IRFs. 'IRFS and FE Only' impose the VAR restrictions as described in the text. '8 lags' means that a VAR with eight lags was used instead of a VAR with four lags.

and Wolf (2021). Table A.1 reports empirical rejection rates for different versions of the robust SP-IV test statistics for simulations using the true monetary policy shock as the instrument. The "FE only" version uses the forecast errors implied by the VAR but does not impose the VAR restrictions on the IRFs. The "IRFs and FE" version additionally imposes the VAR restrictions on both the forecast errors and the IRFs. In this case, the denominator in the AR statistic is obtained using the Delta Method. The "IRFs and FE, 8 lags" version does the same, except that it is based on a VAR with eight lags such that the VAR IRFs are consistent for the first eight IRF horizons. We restrict attention to specifications with $H = 8$, as the Delta Method approximation quickly becomes computationally costly as H increases.

The first row in the table shows that misspecification in the IRFs due to lag truncation in the VAR does not lead to meaningful size distortions in the AR test if the VAR is only used to generate the forecast errors. In contrast, the second row shows that lag truncation bias in the IRFs at longer horizons creates size distortions when the VAR restrictions are also imposed on the IRFs. The distortions become larger as sampling error fades with larger T, and when a larger number of biased IRF estimates

are used for identification as N_z increases. The third row shows that, when T is large, increasing the lag length in the VAR to eight essentially eliminates the size distortions. However, including more lags in the VAR creates large distortions in small samples because of the larger number of VAR parameters to be estimated. In practice, researchers concerned with VAR lag truncation bias can choose lag length as a function of sample size, or else simply not impose the VAR restrictions on the IRFs when using the AR test for inference. In the simulations in Section 3, we chose the latter option and report the "FE only" AR test while keeping the VAR lag length at four for all sample sizes.

Finally, the next three rows in Table A.1 show that our implementation of the KLM test for SP-IV is not affected by lag truncation bias. All empirical rejection rates remain close to nominal size regardless of T, N_z, the number of lags in the VAR, or whether the VAR restrictions are imposed on the IRFs or not. Given the greater robustness in the simulations, we recommend inference based on the KLM test when using the VAR implementation for SP-IV. In the simulations in Section 3 as well as the empirical application in Section 4, we use the "IRFs and FE" version of the KLM test that imposes the restrictions implied by a VAR on the IRFs.

B. Generalized and CUE SP-IV

Using the weighting matrix $\Phi_s(\beta,\zeta) = (\Sigma^{-1}_{u^\perp_H} \otimes Q^{-1})$, where $\Sigma_{u^\perp_H}$ is the covariance of $u^\perp_{H,t}$, leads to the efficient GMM estimator of β. This estimator is also the "Generalized Least Squares" version of SP-IV minimizing $\text{Tr}\left((u^\perp_H P_{Z^\perp} u^{\perp\prime}_H)\Sigma^{-1}_{u^\perp_H}\right)$. Given $\Sigma_{u^\perp_H}$, the closed form generalized SP-IV estimator is

(B.1)
$$\hat{\beta}_G = \left(R'\left(Y^\perp_H P_{Z^\perp} Y^{\perp\prime}_H \otimes \Sigma^{-1}_{u^\perp_H}\right)R\right)^{-1} R'\left(Y^\perp_H P_{Z^\perp} \otimes \Sigma^{-1}_{u^\perp_H}\right)\text{vec}(y^\perp_H P_{Z^\perp}) \ .$$

For inference, we replace Assumption 2.d by

Assumption 2.d′. *$R'(\Theta_Y\Theta'_Y \otimes \Sigma^{-1}_{u^\perp_H})R$ is a fixed matrix with full rank.*

Under Assumptions 1, 2.a-2.c, Assumption 2.d′, and Assumption 3,

$$\text{(B.2)} \quad \sqrt{T}(\hat{\beta}_G - \beta) \xrightarrow{d} N(0, V_{\beta_G}) \quad , \quad V_{\beta_G} = \left(R' \left(\Theta_Y \Theta_Y' \otimes \Sigma_{u_H^\perp}^{-1} \right) R \right)^{-1}.$$

The Generalized SP-IV estimator is feasible after replacing $\Sigma_{u_H^\perp}$ with a consistent estimator like the one in Section 2.1, using a two-step or iterated procedure. Alternatively, the continuously updating (CUE) GMM estimator minimizes the AR statistic in (20) with respect to b. The KLM statistic in (21) is zero at the CUE estimator, so both AR and KLM confidence sets contain the CUE.

C. The Impact of Estimation Error on Inference

As noted in the text, after plugging in estimated forecast errors, it is not necessary to conduct any explicit adjustment for estimation error in $\hat{\zeta}$ for asymptotically valid inference on $\hat{\beta}$. This follows from the standard asymptotic variance of the GMM estimator implicitly defined by the objective function (7). However, to make the reasoning explicit, denote the forecasting model as $E[y_{H,t}|X_{t-1}] = g_1(\zeta)X_{t-1}, E[Y_{H,t}|X_{t-1}] = g_2(\zeta)X_{t-1}, E[z_t|X_{t-1}] = g_3(\zeta_t)X_{t-1}$, which nests both LPs and VARs. Computing the expected Jacobian of the moments in (5) with respect to d at $d = \zeta$, $b = \beta$ yields

$$\text{(C.3)} \quad E\left[\frac{\partial z_t^\perp \otimes u_{H,t}^\perp}{\partial \zeta'}\right] = E\left[-\frac{\partial g_3(\zeta)}{\partial \zeta'} X_{t-1} \otimes u_{H,t}^\perp + z_t^\perp \otimes \left(-\frac{\partial g_1(\zeta)}{\partial \zeta'} X_{t-1} + \beta \frac{\partial g_2(\zeta)}{\partial \zeta'} X_{t-1}\right)\right]$$

$$\text{(C.4)} \quad = -\frac{\partial g_3(\zeta)}{\partial \zeta'} E\left[X_{t-1} \otimes u_{H,t}^\perp\right] + E\left[z_t^\perp \otimes \left(-\frac{\partial g_1(\zeta)}{\partial \zeta'} + \beta \frac{\partial g_2(\zeta)}{\partial \zeta'}\right) X_{t-1}\right]$$

$$\text{(C.5)} \quad = 0,$$

where the last equality follows because $u_{H,t}^\perp$ and $z_t^\perp$ are orthogonal to X_{t-1} by construction. The Jacobian of (6) with respect to β is likewise zero since the forecasting moments do not depend on β. Therefore, the expected Jacobian, $J(\beta, \zeta)$, of $f(\cdot\,; \beta, \zeta)$ is block diagonal. If $\Phi(b, d)$, the weighting matrix, is also block diagonal, then $J'\Phi J$ is block diagonal, as

is its inverse. Let V_f denote the variance-covariance of $f(\cdot\,;\beta,\zeta)$. Then, in the asymptotic variance of $(\beta',\zeta')'$, $(J'\Phi J)^{-1}J'\Phi V_f \Phi J(J'\Phi J)^{-1}$, each diagonal block depends only on the corresponding diagonal block of V_f. Therefore, plugging in the estimated forecast errors into the moments $f_s(\cdot\,;\beta)$ and henceforth treating them as data and using the standard formula based on just those moments yields a valid asymptotic variance, without further adjustment. Alternatively, a delta method argument can be applied directly to the sample counterpart of (5) to show that its asymptotic variance does not depend on estimation error in $\hat{\zeta}$ due to (C.5).

D. Proof of Proposition 2

Proof. The asymptotic variance of the SP-IV estimator in (9) is

$$aVar(\hat{\beta}) = (\boldsymbol{\Theta}_Y'\boldsymbol{\Theta}_Y)^{-1}\boldsymbol{\Theta}_Y'\left(I_{N_z}\otimes \mathrm{var}(u_{H,t}^{\perp})\right)\boldsymbol{\Theta}_Y(\boldsymbol{\Theta}_Y'\boldsymbol{\Theta}_Y)^{-1}\,, \tag{D.1}$$

The asymptotic variance of the 2SLS estimator is

$$aVar(\hat{\beta}_{2SLS}) = (\boldsymbol{\Theta}_Y'\boldsymbol{\Theta}_Y)^{-1}\boldsymbol{\Theta}_Y'\Omega(z_{H,t}u_t)\boldsymbol{\Theta}_Y(\boldsymbol{\Theta}_Y'\boldsymbol{\Theta}_Y)^{-1}\,, \tag{D.2}$$

where $\Omega(z_{H,t}u_t)$ is the long-run variance of $z_{H,t}u_t$, and $z_{H,t}$ stacks the N_z $H\times 1$ vectors $z_{H,it}$ of lags $0,\ldots,H-1$ of z_{it}. Without loss of generality, assume z_t is normalized to have identity covariance. Because z_t is i.i.d., only the first $H-1$ autocovariances of $z_{H,t}$ are non-zero, mechanically due to the overlapping lag structure. For lag l, $|l|<H$, $\mathrm{cov}(z_{H,it}, z_{H,i(t-l)}) = \iota_{-l}$, where ι_{-l} is the $-l$-diagonal matrix (i.e., a matrix of zeros except for ones along the diagonal l entries below the main diagonal), and $\mathrm{cov}(z_{H,t}, z_{H,t-l}) = I_{N_z}\otimes\iota_{-l}$. It follows that

$$\mathrm{cov}(z_{H,t}u_t, z_{H,t-l}'u_{t-l}) = (I_{N_z}\otimes\iota_{-l})\gamma_l,\ |l|<H, \tag{D.3}$$

and zero otherwise, where γ_l is the l^{th} autocovariance of u_t. Then

$$\Omega(z_{H,t}u_t) = \sum_{l=-\infty}^{\infty}\mathrm{cov}(z_{H,t}u_t, z_{H,t-l}'u_{t-l}) = \sum_{l=-H+1}^{H-1}(I_{N_z}\otimes\iota_{-l})\gamma_l. \tag{D.4}$$

The last sum is block-diagonal, and under stationarity each $H \times H$ block is equal to the matrix with γ_l along the l^{th} diagonal, which is the auto-covoariance matrix of u_t (up to lag $H-1$), or Σ_{u_H}. If u_t is i.i.d., or if X_{t-1} is an irrelevant predictor or just a constant, $u_{t+h} = u_t(h) = u_t^{\perp}(h)$, so $\Sigma_{u_H} = \Sigma_{u_H^{\perp}}$, and $\Omega(z_{H,t}u_t) = I_{N_z} \otimes \Sigma_{u_H^{\perp}}$, establishing (i).

For (ii), we consider $\hat{\beta}_j$ asymptotically more efficient than $\hat{\beta}_i$ if $aVar(\hat{\beta}_i) - aVar(\hat{\beta}_j)$ is positive semi-definite (Rothenberg and Leenders 1964). $aVar(\hat{\beta}_{2SLS}) - aVar(\hat{\beta})$ is positive definite if maxeval $(\Sigma_{u_H}) >$ maxeval $\left(\Sigma_{u_H^{\perp}}\right)$.

□

E. Proof of Proposition 3

Proof. Consider the weak instruments asymptotic embedding $\Theta_Y = C/\sqrt{T}$ where C is a $HK \times N_z$ fixed matrix. When $K = 1$, the concentration parameter for 2SLS is $\text{Tr}(CC')/\text{Tr}(\Omega(z_{H,t}\omega_t)) = \text{Tr}(CC')/(HN_z\sigma_\omega^2)$, where σ_ω^2 is the variance of the first stage error term, since the long-run variance of $z_{H,t}\omega_t$ has the same structure as that of $z_{H,t}u_t$ above (see definitions in Lewis and Mertens (2022) and Montiel-Olea and Pflueger (2013)). For SP-IV without projecting onto X_{t-1}, the concentration parameter is $\text{Tr}(CC')/(N_z \text{Tr}(\Sigma_{v_H}))$, see Definition 1 in the Online Appendix. For SP-IV with conditioning on X_{t-1}, the concentration parameter is $\text{Tr}(CC')/(N_z \text{Tr}(\Sigma_{v_H^{\perp}}))$, see Definition 1 in the Online Appendix. $\text{Tr}(\Sigma_{v_H})$ is larger than $\text{Tr}(\Sigma_{v_H^{\perp}})$ unless X_{t-1} is completely irrelevant for predicting $Y_{t+h}, h = 0, ..., H-1$. Parts (i) and (ii) follow from the expressions for the concentration parameters.

□