Quantitative Economics 14 (2023), 161–200 1759-7331/20230161

Monetary policy, external instruments, and heteroskedasticity

Thore Schlaak
Department of Economics, Freie Universität Berlin

Malte Rieth
Department of Economics, Martin-Luther-Universität Halle-Wittenberg and DIW Berlin

Maximilian Podstawski
DIW Berlin

We develop a structural vector autoregressive framework that combines external instruments and heteroskedasticity for identification of monetary policy shocks. We show that exploiting both types of information sharpens structural inference, allows testing the relevance and exogeneity condition for instruments separately using likelihood ratio tests, and facilitates the economic interpretation of the structural shock of interest. We test alternative instruments and find that narrative and model-based measures are valid, while high-frequency data instruments show signs of invalidity. Finally, we document that monetary shocks identified with both a valid instrument and heteroskedasticity have larger effects on production and prices than monetary shocks identified via an instrument only.

Keywords. Monetary policy, structural vector autoregressions, identification with external instrument, heteroskedasticity, Markov switching.

JEL classification. C32, E32, E52, E58.

1. Introduction

Estimating the effects of monetary policy is a central element of macroeconomic analysis. While the economy reacts to policy decisions, monetary policy is also endogenous to the state of the economy, posing the issue of isolating exogenous variation in monetary policy. In the empirical literature, structural vector autoregressions (SVARs) are a main tool for studying the causal effects of monetary interventions. Departing from the classical identification via zero restrictions, two identification approaches are receiving increasing attention in the literature. On the one hand, authors use external data

Thore Schlaak: schlaak@zedat.fu-berlin.de
Malte Rieth: mrieth@diw.de
Maximilian Podstawski: maximilian.podstawski@gmail.com

We are thankful to three anonymous referees, Nadav Ben Zeev, Dario Caldara, Lutz Kilian, Alexander Kriwoluzky, Helmut Lütkepohl, Geert Peersman, Dieter Nautz, Konstantinos Theodoridis, Giovanni Ricco, Mathias Trabandt, Lars Winkelmann, Rafael Wouters, as well as the conference and seminar participants at the Workshop on Structural VAR Topics Queen Mary University London, Freie Universität Berlin, DIW Berlin, Verein für Socialpolitik, Royal Economic Society Annual Conference, European Economic Association Annual Conference, and International Association of Applied Econometric Annual Conference.

 https://doi.org/10.3982/QE1511

on monetary surprises to identify latent monetary shocks in SVARs.[1] On the other hand, many papers draw on volatility changes in macroeconomic and financial data to identify monetary shocks.[2] Both strategies are popular because they are parsimonious in terms of identifying assumptions and because they incorporate further information into the model.

Identification via an external instrument allows for a contemporaneous response of monetary policy to asset prices. Moreover, it adds a potentially large information set to the model through a narrative or financial data-based instrument. Finally, it accounts for measurement error in the instrument, which reduces the attenuation bias in models treating the proxy as the true shock (Mertens and Ravn (2013)). However, these advantages rely on the presumption that the instrument is valid, that is, relevant and exogenous.

Identification through heteroskedasticity adds information from time-varying second moments to the model and relies on even weaker identifying assumptions. While an instrument for monetary policy shocks needs to move interest rates without correlating with other structural shocks, a relative increase in the variance of monetary shocks can be sufficient to trace out the response of the other variables in the system to these shocks. The relative variance shift can be viewed a "probabilistic instrument" that increases the likelihood that monetary policy shocks occur (Rigobon (2003)). Again, these minimal assumptions are not costless. The statistically identified shocks are often economically difficult to interpret.

This paper proposes a framework that combines both identification strategies to improve inference within SVARs. The framework preserves the attractive features of both approaches but addresses some of the key limitations that each of them has in isolation. It makes use of external instruments for monetary policy shocks proposed in the literature. In addition, it exploits time-variation in the second moments of the data. The combination of both types of identifying information into a "heteroskedastic proxy-SVAR" has three main advantages relative to models using only one type of information.

First, the encompassing framework sharpens the identification of the structural model, and hence the suitability of the model for policy analysis. To show this, we conduct an extensive simulation study. The Monte Carlo evidence suggests that the encompassing model yields more accurate estimates of the true parameters according to the mean squared errors of impulse response functions than either of the two identification approaches in isolation.

The second advantage of our framework is that it allows testing the validity, that is, relevance and exogeneity, of external instruments. We include the instrument as an endogenous variable in an augmented SVAR, as in Caldara and Herbst (2019). As changes in volatility of the residuals of the augmented model can suffice for point-identification of the full structural model, additional restrictions, and hence the exogeneity condition

[1] See Gertler and Karadi (2015), Cesa-Bianchi, Thwaites, and Vicondoa (2016), Miranda-Agrippino and Ricco (2018), Stock and Watson (2018), Rogers, Scotti, and Wright (2018), and Caldara and Herbst (2019).

[2] See Rigobon and Sack (2004), Normandin and Phaneuf (2004), Lanne and Lütkepohl (2008), Wright (2012), Herwartz and Lütkepohl (2014), and Nakamura and Steinsson (2018).

becomes testable. This conveniently reduces to testing zero restrictions on the structural impact matrix of the augmented SVAR. We propose a likelihood ratio (LR) test for that purpose. Monte Carlo evidence suggests that it has desirable properties in terms of size and power. Testing the exogeneity condition has so far been unresolved in the literature but is of particular interest as the violation of this condition may lead to erroneous conclusions regarding the validity of the instrument and the effects of latent structural shocks. We also propose a LR-test for evaluating the relevance condition. The Monte Carlo evidence suggests that the test reliably discriminates between relevant and irrelevant instruments. It thereby complements existing versions of F-tests for instrument relevance (Stock, Wright, and Yogo (2002), Stock and Watson (2012), Mertens and Ravn (2013)). In our set-up, it has more power than the F-test because it uses all information in the model both under the null and the alternative hypothesis.

The third advantage of the framework addresses a main challenge in the literature on identification through heteroskedasticity (Rigobon and Sack (2003), Herwartz and Lütkepohl (2014)). In this class of models, structural shocks are identified statistically. They need to be labeled by the researcher after estimation. While the literature has developed several strategies for that purpose, this task is often difficult and can cast doubt on the economic meaning of the structural shocks. Our framework simplifies the interpretation of the structural shocks, because the inclusion of a relevant instrument based on prior economic reasoning into the model pins down the shock of interest.

We use our framework to evaluate the validity of instruments proposed in the literature and to provide new, and in light of the Monte Carlo evidence, sharper estimates of the macroeconomic effects of monetary policy shocks in the United States based on heteroskedasticity and a valid instrument. The former is a common and well-documented feature of U.S. real and financial data (Stock and Watson (2002), Justiniano and Primiceri (2008), Amir-Ahmadi, Matthes, and Wang (2016)). Standard statistics provide strong evidence that changes in volatility are also present in our sample. We model them within a Markov switching in variances framework and use them for identification. For the latter, we include the measure of unanticipated changes in the intended federal funds rate of Romer and Romer (2004). This measure is a cornerstone of monetary policy analysis but also criticized for being predictable or endogenous (Leeper (1997), Miranda-Agrippino and Ricco (2018)). We use our LR-tests to evaluate these claims. The evidence suggests that the instrument is contemporaneously exogenous to demand, supply, and cost-push shocks, supporting the approach of Romer and Romer (2004) and studies using it as an instrument for monetary policy shocks (Stock and Watson (2012), Tenreyro and Thwaites (2016), Rey (2016)). We find that an unexpected increase in the federal funds rate by 100 basis points leads to a fall of economic activity by 1.6% and of consumer prices by 0.8%. These effects are twice as large as estimates obtained from a proxy-SVAR that does not exploit heteroskedasticity.

We also test and compare alternative instruments for monetary shocks. We find that model-based measures (Bernanke, Boivin, and Eliasz (2005)) are also exogenous, while higher frequency instruments show signs of invalidity (Barakchian and Crowe (2013), Gertler and Karadi (2015), Miranda-Agrippino and Ricco (2018)). Nevertheless,

the models including high-frequency instruments all imply a significant decline in output and prices in response to an unexpected tightening. This finding illustrates another advantage of exploiting the time-varying volatility in proxy-SVARs. The heteroskedasticity adds sufficient information to identify the monetary shocks, while it allows dealing with weak instruments. If there is sufficient time-variation in the second moments—a condition that can be checked empirically after estimation—the model is statistically identified even with little identifying information from the external instrument. In such a situation, the relevance condition is no longer necessary for reliable inference. Instead, it reduces to a question about the information contained in the instrument and the interpretation of the associated structural shock.

This paper is related to several recent articles on the identification of SVARs. Bertsche and Braun (2020) propose using stochastic volatility for identification of SVARs. They focus on the econometric theory of using a stochastic volatility model and apply their setup to the oil market. After estimation, they project the estimated structural shocks onto candidate instruments in a second step. For that auxiliary regression, they derive Wald-type tests for instrument validity outside the SVAR. This is different to our approach, which tests the instruments within the SVAR and uses LR-tests. Moreover, they do not use the instrument as additional source of identifying information in an encompassing model. In this respect, our analysis is more closely related to Antolín-Díaz and Rubio-Ramírez (2018) who develop a framework that exploits two types of identifying information to improve inference in SVARs. They combine sign restrictions with prior information on specific shocks and use a Bayesian setting, whereas we combine heteroskedasticity with instruments in a frequentist approach. Ludvigson, Ma, and Ng (2017) propose an identification strategy of SVARs through prior knowledge of certain shocks within classical inference. They show how the resulting new type of inequality restrictions can be combined with external instruments to further sharpen inference. Finally, Arias, Rubio-Ramirez, and Waggoner (2021) develop algorithms for exact finite sample inference in Bayesian proxy-SVAR models. Their framework is sufficiently flexible to allow for multiple instruments. However, the latter two articles are not concerned with testing the validity of the instruments.

The remainder of the paper is structured as follows. The next section introduces the heteroskedastic proxy-SVAR and discusses identification, testing, and estimation. Section 3 presents simulation results in support of the framework. In Section 4, we use the heteroskedastic proxy-SVAR to shed new light on the efficacy of monetary policy and to test a range of instruments discussed in the literature. Finally, Section 5 concludes. The Online Appendix may be found within the replication file (Schlaak, Rieth, and Podstawski (2023)).

2. The SVAR framework

The vector autoregressive (VAR) model is

$$y_t = \gamma + A(L)y_{t-1} + u_t, \tag{1}$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ is a $(K \times 1)$-vector of observable variables, $A(L)$ is a lag matrix polynomial capturing the autoregressive component of the model, γ collects constant terms, and the u_t are K-dimensional serially uncorrelated observable residuals. The residuals u_t are linearly related to white noise structural shocks ε_t according to

$$u_t = B\varepsilon_t. \tag{2}$$

We assume that the VAR is invertible and has a Wold moving average representation $y_t = \alpha + \sum_{i=0}^{\infty} \Phi_i u_{t-i}$.

2.1 *A heteroskedastic proxy-SVAR*

A common feature of macroeconomic and financial data are changes in volatility over time (see, among others, Stock and Watson (2002), Justiniano and Primiceri (2008), Amir-Ahmadi, Matthes, and Wang (2016)). Rigobon and Sack (2004), Normandin and Phaneuf (2004), and Lanne and Lütkepohl (2008) show that this holds in particular for the analysis of monetary policy where changes in volatility of the data feed into heteroskedastic residuals in monetary SVARs and can be used for identification. Against this backdrop, we allow for heteroskedastic residuals in (1).[3] We assume that the volatility changes are driven by a first-order Markov switching (MS) process $S_t \in \{1, \ldots, M\}$ with M states and transition probabilities $p_{kl} = P(S_t = l | S_{t-1} = k)$, $k, l = 1, \ldots, M$. Furthermore, the reduced form residuals are normally and independently distributed conditional on a given state $u_t | S_t \sim \text{NID}(0, \Sigma(S_t))$, where all Σ_m, $m = 1, \ldots, M$ are distinct.

Another prominent way to identify structural shocks is via external instruments (Stock and Watson (2012), Mertens and Ravn (2013)). We assume that the process generating the (potentially heteroskedastic) instrument w_t has a linear form, following Caldara and Herbst (2019):

$$w_t = \beta\varepsilon_t + \eta\nu_t, \tag{3}$$

where ε_t is the $K \times 1$ vector of structural shocks, $\beta = (\beta_1, \ldots, \beta_K)$ is a $1 \times K$-coefficient vector, $\nu_t \sim N(0, \sigma_m^2)$ is a measurement error uncorrelated with the structural shocks ε_t, and η scales the effect of the noise. Without loss of generality, we order the structural shock of interest first. Then β_1 and η may be interpreted as weighting parameters of signal to noise.

Instrument validity requires the following two conditions:

$$\beta_i = 0 \quad \forall i = 2, \ldots, K, \tag{4}$$

$$\beta_1 \neq 0. \tag{5}$$

[3] We refrain from introducing additional nonlinearity into the model by allowing state dependency in the constant or autoregressive parameters as we are interested in the heteroskedasticity features of the data for identification purposes.

Equation (4) is the exogeneity and (5) the relevance condition. If both are met, the covariances between the instrument, the shock of interest, and the other shocks are

$$\operatorname{Cov}(w_t, \varepsilon_{1,t}) = \mathbb{E}[w_t \varepsilon_{1,t}] = \mathbb{E}\left[\beta_1 \varepsilon_{1,t}^2 + \eta \nu_t \varepsilon_{1,t}\right] = \beta_1 \mathbb{E}\left[\varepsilon_{1,t}^2\right] = \beta_1 \operatorname{Var}(\varepsilon_{1,t}) \neq 0, \tag{6}$$

$$\operatorname{Cov}(w_t, \varepsilon_{i,t}) = \mathbb{E}[w_t \varepsilon_{i,t}] = \mathbb{E}[\beta_1 \varepsilon_{1,t} \varepsilon_{i,t} + \eta \nu_t \varepsilon_{i,t}] = 0 \quad \forall i = 2, \ldots, K, \tag{7}$$

which use (4), (5), and the independence of ε_t and ν_t. Equation (6) shows that a valid instrument is only related to the shock of interest through β_1. If the variance of that shock changes across the regimes, so will the covariance. The assumption of constant β across regimes thus attributes potential changes in the covariance to heteroskedasticity in $\varepsilon_{1,t}$. Equation (7) shows that the covariance of the instrument with the other shocks is zero even if their variances change.[4]

We compile the system by appending model (1) with (3). The augmented VAR is

$$z_t = \delta + \Gamma(L) z_{t-1} + e_t, \tag{8}$$

where $z_t = [y_t', w_t]'$ is a $((K+1) \times 1)$-vector of observable variables, $\Gamma(L)$ is a lag matrix polynomial capturing the autoregressive component of the model, δ is a $((K+1) \times 1)$-vector of constant terms, and e_t are $(K+1)$-dimensional serially uncorrelated residuals.[5] The latter are related to the structural innovations μ_t as

$$\begin{aligned} e_t &= D \mu_t \\ &= \begin{bmatrix} B_{(K \times K)} & 0_{(K \times 1)} \\ \beta_{(1 \times K)} & \eta \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \nu_t \end{bmatrix}. \end{aligned} \tag{9}$$

Using (9), we rewrite the augmented VAR in (8) in structural form as

$$z_t = \delta + \Gamma(L) z_{t-1} + D \mu_t. \tag{10}$$

Since the state dependency in the variances of the reduced form residuals in (8), $\operatorname{var}(e_t | m) = \tilde{\Sigma}_m$ with $m = 1, \ldots, M$, translates into the structural form, we have $E[\mu_t] = 0$ and $E[\mu_t \mu_t' | m] = \Lambda_m$, where Λ_m is a diagonal matrix satisfying the orthogonality condition of the structural innovations. This heteroskedasticity pattern provides a valuable source of identifying information (Rigobon and Sack (2004), Normandin and Phaneuf (2004), Lanne and Lütkepohl (2008)). Under the assumption of a constant instantaneous impact matrix D, for each volatility regime a decomposition

$$\tilde{\Sigma}_m = D \Lambda_m D' \tag{11}$$

exists, where $\Lambda_m = \operatorname{diag}(\lambda_{1,m}, \ldots, \lambda_{K+1,m})$. We normalize $\Lambda_1 = I_{K+1}$. For $m \geq 2$, the Λ_m are diagonal matrices with positive elements that can be interpreted as the changes of the structural variances in the respective regime relative to the first regime.

[4]The set-up may be extended to multiple instruments. We consider one extension with two instruments for the identification of one structural shock in the simulation study in Section 3.

[5]$\Gamma(L)$ can be restricted if the instrument should not be cleansed, that is, regressed on the lags of other endogenous variables. We do this in the empirical application to work with the raw instrument data and provide a clean comparison to the literature.

Lanne, Lütkepohl, and Maciejowska (2010) state conditions for local uniqueness of matrix D. Local uniqueness implies that D is identified up to the signs of the parameters in each column as well as to column permutations. The conditions for local uniqueness of D are: (i) the structural impact matrix D is time-invariant; (ii) the structural innovations μ_t are orthogonal; and (iii) there are sufficiently many and distinct changes in the variances of the structural innovations, that is, $\lambda_{km} \neq \lambda_{lm}$ for $k, l \in \{1, \ldots, K+1\}$ with $k \neq l$, $\exists m \in \{2, \ldots, M\}$. The first assumption is standard in SVARs identified with external instruments.[6] The second assumption is common in structural VAR analysis more generally. The third assumption can be checked after estimation by comparing the estimated variances λ_{lm}, with $l = 1, \ldots, K+1$.

To see how the reduced form and the structural model are related and how the instrument helps identify the structural parameters, note that

$$\tilde{\Sigma}_m = D\Lambda_m D' = \begin{bmatrix} & & & \beta\Lambda_m b_{.1} \\ & B\Lambda_m B' & & \beta\Lambda_m b_{.2} \\ & & & \vdots \\ & & & \beta\Lambda_m b_{.K} \\ \beta\Lambda_m b_{.1} & \beta\Lambda_m b_{.2} & \ldots & \beta\Lambda_m b_{.K} \quad \beta\Lambda_m\beta' + \eta^2\sigma_m^2 \end{bmatrix}, \tag{12}$$

where $b_{.j}$ denotes the jth column of B. The last column (or row) of (12) summarizes the restrictions on the structural parameters of the model implied by the instrument. If the instrument is valid, the K first elements of that column satisfy $\mathbb{E}[u_t w_t] = \mathbb{E}[w_t(b_{.1}\varepsilon_{1t} + B^*\varepsilon_t^*)] = b_{.1}\mathbb{E}[w_t\varepsilon_{1t}] = b_{.1}\beta_1\lambda_1$, where B^* contains the $2, \ldots, K$ remaining columns of B and ε_t^* the corresponding structural shocks. The expression $b_{.1}\beta_1\lambda_1$ shows how a valid instrument informs estimation about the first column of B, which is the vector of interest. Moreover, it follows that $\frac{\mathbb{E}[u_{it}w_t]}{\mathbb{E}[u_{1t}w_t]} = \frac{b_{i1}}{b_{11}}$. This ratio shows that the relative impact responses are not affected by changes in volatility.

2.2 *Testing the validity of an instrument*

While our approach is novel in combining information of an instrument and heteroskedasticity for identification, it conveniently reduces to the standard case for identification via heteroskedasticity as it incorporates the instrument into the augmented SVAR (see (10)). Hence, estimation, identification, and testing follows Lanne and Lütkepohl (2008) and Lanne, Lütkepohl, and Maciejowska (2010). Specifically, if the conditions for local uniqueness are met, the heteroskedasticity in the residuals allows for estimating all structural parameters of D. Any additional restrictions on D are then overidentifying, and hence testable. This is particularly interesting in our context as it allows for testing both the relevance and the exogeneity of the instrument, and thus its validity. Such tests reduce to testing zero restrictions on β, that is, the last row of the structural impact matrix D. This may be done with likelihood ratio tests (LR-tests) because the elements of β are fixed parameters under the null hypothesis. This implies that the

[6] See Stock and Watson (2012), Mertens and Ravn (2013), Gertler and Karadi (2015), Miranda-Agrippino and Ricco (2018), or Caldara and Herbst (2019).

distribution of the LR-tests is χ^2 with the degrees of freedom equal to the number of restrictions.

To test the exogeneity condition (4), we compare the likelihood of an appropriately restricted version of model (10), that is, we restrict $\beta = (\beta_1, 0, \ldots, 0)$, with an unrestricted version where $\beta = (\beta_1, \beta_2, \ldots, \beta_K)$. Formally, we test

$$H_0 : \beta_2 = \cdots = \beta_K = 0,$$
$$H_1 : \exists j \in \{2, \ldots, K\} \quad \text{s.t. } \beta_j \neq 0.$$

Rejecting the null indicates endogeneity of the instrument.

To test the relevance condition (5), we compare a restricted version of model (10), where $\beta_1 = 0$, to model (10) with β_1 unrestricted. Under both the null and the alternative hypothesis, $\beta_2 = \cdots = \beta_K = 0$. Formally, we test

$$H_0 : \beta_1 = 0,$$
$$H_1 : \beta_1 \neq 0.$$

Rejecting the null indicates the relevance of the instrument.[7] If the instrument is both exogenous and relevant, it is valid. Then we set $\beta = (\beta_1, 0, \ldots, 0)$ and refer to model (10) as a "heteroskedastic proxy-SVAR."

Local uniqueness in our setup implies that a priori we cannot identify the column of the impact matrix D that belongs to a certain structural shock. Practically, this is of little concern. First, assessing the exogeneity of an instrument does not require a particular ordering of the structural shocks. The test will reject the null of all but one β-element equal to zero in case of endogeneity. Second, because an exogenous instrument imposes additional restrictions on the covariance matrix, the shock that is most consistent with these restrictions will be ordered to the column with the only unrestricted element of β. This pins down the shock of interest. In case of an uninformative or weak instrument, the structural shock of interest would not necessarily be related to the unrestricted element of β. However, such a situation would not affect inference. It would only dilute the economic interpretation of the results. The relevance test would indicate such situations as uninformative instruments will be detected through not rejecting the null. Then it is up to the researcher whether these should be included nevertheless, or discarded.

There are two related papers, which assess the validity of external instruments for overidentified SVARs. Cesa-Bianchi, Thwaites, and Vicondoa (2016) propose an auxiliary GMM estimation and a Hansen–Sargan statistic to test the validity of their baseline instrument with a second external instrument. Angelini and Fanelli (2018) provide a general framework for employing multiple instruments for shock identification and show how to make use of the resulting overidentifying restrictions for a specification test of the proxy-SVAR. Our tests complement these approaches but differ along an important dimension. While these papers use multiple external instruments, which are potentially subject to the same endogeneity concerns, we employ structural shocks identified

[7] Alternatively, one can test for instrument relevance by testing the null hypothesis that $\beta = 0$ against the alternative that β is unrestricted, with the degrees of freedom equal to the number of columns of β. This test does not assume exogeneity.

through heteroskedasticity for that purpose, which are by construction orthogonal to each other (but do not need to have an economic interpretation).

2.3 *Estimation and bootstrapping*

The parameters of model (10) are estimated by means of the expectation maximization (EM) algorithm proposed by Herwartz and Lütkepohl (2014). Crucial for the analysis is to incorporate the regime-switching nature of the covariance matrix described in (11), given the restrictions on D and Λ_m. All other parameters are assumed to be regime-independent and do not vary across states. For computational details of the EM algorithm, we refer to Section A.1 of the Online Appendix.

Standard errors of the point estimates of the model parameters are obtained from the inverse of the negative Hessian matrix evaluated at the optimum after convergence of the EM algorithm. We use the standard errors of the elements of Λ_m to construct confidence intervals around the point estimates to determine whether they differ significantly from each other. This is a requirement for statistical identification, and hence for the over-identification tests to have sufficient power.[8]

For inference on the structural impulse response functions, bootstrapped pointwise confidence bands are computed. Given the heteroskedastic pattern of the data, a simple reshuffling of the estimated residuals $\hat{e}_t$, as in a classic residual bootstrap, does not preserve the second moment properties of the data and invalidates inference. Hence, we use a recursive design wild bootstrap and construct bootstrapped samples as

$$z_t^* = \hat{\delta} + \hat{\Gamma}(L) z_{t-1} + \varphi_t \hat{e}_t,$$

where $\hat{\delta}$ and $\hat{\Gamma}(L)$ are estimated counterparts of the coefficients defined in (10), and φ_t is an independent random variable following a Rademacher distribution, that is, φ_t is either 1 or -1 with probability 0.5. Each of the 1000 generated bootstrap samples is based on identical presample values from the original data set as initial values, that is, $z^*_{-p+1} = z_{-p+1}, \ldots, z^*_0 = z_0$. The bootstrap is conducted conditionally on estimated parameters for the relative variances and transition probabilities, which is a commonly used technique for these types of models (Herwartz and Lütkepohl (2014), Podstawski and Velinov (2018)). Our results are robust to using two related bootstrap procedures. First, we explore a residual-based moving block bootstrap proposed for proxy-VARs by Jentsch and Lunsford (2016). Second, we use draws from a normal distribution in a fixed-design recursive bootstrap. Lütkepohl and Schlaak (2019) show that this methods perform well for a model with volatility by driven by GARCH. We refer to Online Appendix A.4 for the computational details of the respective bootstraps.

[8]For the class of MS-models, currently no formal statistical tests for identification are available. As the model under the null hypothesis may not be identified, the derivation of the asymptotic distribution of Wald- or LR-tests is not straightforward. For this reason, in the existing literature, usually the point estimates and standard errors of the respective elements of Λ_m are considered when checking for identification (Herwartz and Lütkepohl (2014)).

3. Simulation study

To explore the properties of our framework and LR-tests, we conduct an extensive Monte Carlo study. We evaluate how the tests behave for different degrees of instrument endogeneity and relevance. Then we assess whether the framework improves the accuracy of the estimation of the structural model. We also discuss how the proposed framework simplifies estimation and inference with weak instruments in proxy-SVARs.

3.1 *Setup of Monte Carlo study*

We assume that the data generating process is of the form (10). The process implies that y_t and w_t are jointly normally distributed conditional on state $m = 1, \dots, M$. In the simulation, we generate data for y_t and then for w_t contingent on the realizations of y_t. We use the following parameters of a structural first-order autoregressive model, which are taken from the New Keynesian DSGE-model of An and Schorfheide (2007):

$$\begin{bmatrix} r_t \\ x_t \\ \pi_t \end{bmatrix} = \begin{bmatrix} 0.79 & 0.00 & 0.25 \\ 0.19 & 0.95 & -0.46 \\ 0.12 & 0.00 & 0.62 \end{bmatrix} \begin{bmatrix} r_{t-1} \\ x_{t-1} \\ \pi_{t-1} \end{bmatrix} + \begin{bmatrix} 0.69 & 0.61 & 0 \\ -1.10 & 1.49 & 1 \\ -0.75 & 1.49 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_t^r \\ \varepsilon_t^z \\ \varepsilon_t^g \end{bmatrix},$$

where r_t is the interest rate, x_t is output, and π_t is the inflation rate. The structural shocks are a monetary policy shock (ε_t^r), a productivity shock (ε_t^z), and a government spending shock (ε_t^g).

The variances of the structural innovations are driven by a discrete MS process with $M = 2$ states and transition probabilities

$$P = \begin{bmatrix} 0.975 & 0.025 \\ 0.050 & 0.950 \end{bmatrix},$$

which are used to generate the Markov states S_t for $t = 1, \dots, T$. Following standard conventions, we normalize the variances of the structural innovations in the first state to unity. We set the relative variances in the second state by choosing rather distinct variances in the range used in comparable studies (Lütkepohl and Schlaak (2018)):

$$\Lambda_2 = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{pmatrix}.$$

Given that the impact matrix is identified up to column signs and permutations only, we assure that the model is uniquely determined by sorting the estimated coefficients of Λ_2 in ascending order and by adjusting the columns of the estimated impact matrix correspondingly. With appropriate starting values $y_0 = (0, 0, 0)'$, we generate data recursively by drawing from

$$\varepsilon_t \sim \begin{cases} N(0, I) & \text{for } m = 1, \\ N(0, \Lambda_2) & \text{for } m = 2, \end{cases}$$

using $B\varepsilon_t = u_t$ to calculate the reduced form residuals.

With the structural innovations at hand, we generate the instrument w_t using (3). We set $\eta = 1$ and the variances of the noise parameter ν_t such that

$$\nu_t \sim \begin{cases} N(0, 1) & \text{for } m = 1, \\ N(0, 12) & \text{for } m = 2. \end{cases}$$

This setup implies a time-varying volatility of the instrument, which can be observed in many time series of instruments that are used in the literature (Romer and Romer (2004), Gertler and Karadi (2015)).

The correlation between the monetary shock and the instruments is determined by β, η, the variance of the monetary shock $\text{Var}_m(\varepsilon_{1,t})$, and the variance of the noise $\text{Var}_m(\nu_t)$. The correlation can change with both variances across states m. The variance of w_t is $\text{Var}_m(w_t) = \beta^2 \text{Var}_m(\varepsilon_{1,t}) + \eta^2 \text{Var}_m(\nu_t)$. Hence, the correlation between $\varepsilon_{1,t}$ and w_t is $Corr_m(\varepsilon_{1,t}, w_t) = \beta\sqrt{\text{Var}_m(\varepsilon_{1,t})}/\sqrt{\beta^2 \text{Var}_m(\varepsilon_{1,t}) + \eta^2 \text{Var}_m(\nu_t)}$.

We set $\beta = (\beta_1, \beta_2, 0)$, where β_1 captures the relevance of the instrument for the monetary shock ε_t^r, while β_2 measures the endogeneity to the second structural shock ε_t^z. We equate β_3 to zero to focus the simulation study, concentrating on cases where endogeneity stems from one source only. We construct different instruments for the monetary policy shock. We target correlations of $\rho_1 \in [0, 0.15, 0.3, 0.4]$ between the monetary shock and the instrument over the whole sample by setting $\beta_1 \in [0, 0.35, 0.72, 1]$. The respective correlations may vary across states. Compared to related studies Lütkepohl and Schlaak (forthcoming), these correlations are small as we are also interested in the case of weak instruments. Similarly, we introduce different degrees of endogeneity by setting $\beta_2 \in [0, 0.05, 0.17, 0.27, 0.37]$ to obtain sample correlations of the instrument with the nonmonetary shock of $\rho_2 \in [0, 0.03, 0.1, 0.15, 0.2]$. Finally, we simulate two sample sizes, $T = 200$ and $T = 500$, which are within the typical range of macroeconomic data sets. The number of replications for each simulation design is $R = 500$.

3.2 *Fitted models*

As a reference model for the test evaluation, we fit a MS(2)-VAR(1) with unrestricted β to the data. Then we estimate and compare the following three models:

Model A Heteroskedastic proxy-SVAR with $\beta = (\beta_1, 0, 0)$, that is, the instrument w_t is assumed to be exogenous.

Model B Heteroskedastic SVAR with $\beta = (0, 0, 0)$, that is, the model is identified via time-varying volatility only.

Model C Standard proxy-SVAR using the identifying information from the external instrument only.

We estimate models A and B as a MS(2)-VAR(1) with respective restrictions on β as discussed in Section 2.1. Model C fits a standard proxy-SVAR with the two stage least squares procedure as suggested by Mertens and Ravn (2013) to evaluate a situation where the volatility in the data is not exploited for identification. Here, the response of the first variable to the identified structural shock is normalized to a have a positive sign. This model has a priori a disadvantage compared to the other models, given

that the generated data feature volatility changes. Alternatively, we estimate a model C* using the methodology proposed by Plagborg-Møller and Wolf (2021), which is robust to invertibility problems. The authors suggest to order the instrument first within a VAR and identify the model through a lower triangular Choleski decomposition. For our setup, their approach is not ideal, however, as it identifies only relative impulse responses, whereas models A and B identify absolute impulse responses. Hence, we can only compare the estimation precision based on the impulse responses unrelated to the impact of the monetary policy shock on the interest rate. In all models, we leave $\Gamma(L)$ unrestricted for a clean comparison.

Given that the reference model and models A and B are nested, we can compute χ^2-distributed LR-statistics to test for the exogeneity and the relevance of the generated instrument in each replication.[9] To test the exogeneity condition, we test the heteroskedastic reference SVAR with β unrestricted against model A. For the relevance condition, we test model A against the more restricted model B.

To assess the benefits of combining identification via external instrument and via heteroskedasticity, for each horizon of the estimated structural impulse response functions for models A–C we calculate the mean squared errors (MSE) and set them into relation to the MSE of model A before cumulating over all impulse response horizons. Thereby, we adjust for scaling differences in the MSE of the individual elements of the response vectors as the absolute magnitude of the impulse responses in the first periods after the shock is typically larger than in later periods given the mean reversion property of the impulse responses. We cumulate the relative MSE for a propagation horizon of up to $h = 25$ such that we capture the impact of differing estimates of both the impact matrix and the autoregressive part of the model since the DGP of our VAR(1) model is fairly persistent. The cumulated relative MSE for horizon h for variable k induced by shock l is calculated as

$$\mathrm{MSE}_h(\theta_{kl,\bullet}) = \frac{1}{h}\sum_{i=0}^{h-1}\left[\left(\frac{1}{R}\sum_{r=1}^{R}\left(\theta_{kl,i} - \hat{\theta}_{kl,i}(r)\right)^2\right) \Big/ \left(\frac{1}{R}\sum_{r=1}^{R}\left(\theta_{kl,i} - \hat{\theta}^{Mod.A}_{kl,i}(r)\right)^2\right)\right], \quad (13)$$

where $\hat{\theta}_{kl,i}(r)$ denotes the estimate of the structural impulse response $\theta_{kl,i}$ of models A–C obtained in the rth replication of our simulation experiment.[10] In what follows, we focus on assessing the accuracy of the parameter estimates associated with the monetary policy shock only.

[9] If $\beta = 0$, the heteroskedastic proxy-SVAR reduces to a standard heteroskedastic SVAR where the distributions of w_t and u_t are independent. In this case, the structural parameters are identified using the heteroskedasticity of the data only. Even if the instrument does not contribute to identification in case of model B, it remains in the autoregressive part of the augmented model and, therefore, may contribute to the estimation of the state probabilities, for example.

[10] The structural impulse responses of the models are obtained as the elements of the matrices $\Theta_i = \Phi_i B$, $i = 0, 1, \ldots$, where Φ_i is the coefficient matrix of the i^{th} propagation horizon of the Wold moving average representation of the VAR. More precisely, the klth element of Θ_i, denoted by $\theta_{kl,i}$, is interpreted as the response of variable k to the lth structural shock after a propagation horizon of i periods.

TABLE 1. Relative rejection frequencies at nominal significance level of 10% of LR-tests on exogeneity of instrument.

		Endogeneity (β_2, ρ_2)				
Sample Size	Relevance (β_1, ρ_1)	(0, 0)	(0.05, 0.03)	(0.17, 0.10)	(0.27, 0.15)	(0.37, 0.20)
$T = 200$	(0, 0)	0.12	–	–	–	–
	(0.35, 0.15)	0.14	0.16	0.54	–	–
	(0.72, 0.30)	0.14	0.17	0.47	0.75	0.93
	(1.00, 0.40)	0.13	0.16	0.40	0.66	0.89
$T = 500$	(0, 0)	0.12	–	–	–	–
	(0.35, 0.15)	0.13	0.22	0.83	–	–
	(0.72, 0.30)	0.12	0.19	0.76	0.99	1.00
	(1.00, 0.40)	0.12	0.17	0.68	0.96	1.00

Note: Each entry in the table is based on 500 replications of each simulation design. Dots (·) denote combinations of values for β_1 and β_2 that produce lower correlations between the instrument w_t and the target structural shock ε_t^r than between the instrument and the nontargeted structural shock ε_t^z. These cases are not taken into account in the analysis.

3.3 *Baseline simulation results*

Table 1 shows the relative rejection frequencies of the LR-test for exogeneity at a nominal significance level of 10% for the two different sample sizes. We focus on this significance level to reduce the possibility of type-II errors, that is, the likelihood of falsely not rejecting an endogenous instrument. The complete set of simulation results, including alternative significance levels, is in Online Appendix A.2. Exogenous instruments with $\rho_2 = 0$ are rejected with relative frequencies reasonably close to their nominal levels (see first column). For $T = 500$, they are very close to the nominal level of 10%.

When moving to the right across columns, the LR-test gains power against the null hypothesis of an exogenous instrument. For both sample sizes, the rejection frequencies steadily increase with higher instrument endogeneity. For $T = 200$, the relative rejection frequencies lie between 40% and 54%, depending on instrument strength, for an endogenous instrument with a correlation of 0.1 with the nonmonetary shock. For $T = 500$ and this correlation, the relative rejection frequencies increase to 68% to 83%. For $\rho_2 \geq 0.15$, endogeneity is reliably detected in at least 66% and often 100% of the cases, depending on the strength of the instrument. For a significance level of 5%, the relative rejection frequencies are slightly lower for small degrees of endogeneity. Higher degrees of endogeneity, ($\rho_2 \geq 0.15$) are reliably detected also at this significance level.

Table 2 displays the relative rejection frequencies of the LR-test for instrument relevance. Now, we focus on a 5% nominal significance level to reduce the type-I error, that is, the probability to accept irrelevant instruments. We use two different significance levels for the exogeneity and relevance tests because we want to be conservative. This approach raises the requirement for instruments to qualify as valid as compared to using a 5% significance level for both tests. For a white noise instrument without any identifying information ($\rho_1 = \rho_2 = 0$), the test shows the expected nominal rejection rate of 5%. When moving south across rows, the rejection frequency rapidly increases for higher correlations of the instrument with the monetary shock. The null of an uninformative

TABLE 2. Relative rejection frequencies at nominal significance level of 5% of LR-test for relevance of instrument.

		Endogeneity (β_2, ρ_2)				
Sample Size	Relevance (β_1, ρ_1)	(0, 0)	(0.05, 0.03)	(0.17, 0.10)	(0.27, 0.15)	(0.37, 0.20)
$T = 200$	(0, 0)	0.05	–	–	–	–
	(0.35, 0.15)	0.94	0.94	0.93	–	–
	(0.72, 0.30)	1.00	1.00	1.00	1.00	1.00
	(1.00, 0.40)	1.00	1.00	1.00	1.00	1.00
$T = 500$	(0,0)	0.05	–	–	–	–
	(0.35, 0.15)	1.00	1.00	1.00	–	–
	(0.72, 0.30)	1.00	1.00	1.00	1.00	1.00
	(1.00, 0.40)	1.00	1.00	1.00	1.00	1.00

Note: Each entry in the table is based on 500 replications of the each simulation design. Dots (·) denote combinations of values for β_1 and β_2 that produce lower correlations between the instrument w_t and the target structural shock ε_t^r than between the instrument and the nontargeted structural shock ε_t^z. These cases are not taken into account in the analysis.

instrument is rejected in all cases and for both sample sizes if $\rho_1 \geq 0.30$, irrespective of the endogeneity.

To obtain an impression of the power of the LR-test and the relevance of the instruments, we compare our test to the well-established F-test for instrument strength (Stock, Wright, and Yogo (2002), Stock and Watson (2012), Mertens and Ravn (2013)). Table 3 contains the relative rejection frequencies at a nominal significance level of 5% and the corresponding F-statistics for exogenous instruments of different strength. The first column shows that the size of the F-test is close to its nominal level for both sample sizes. The rejection frequencies increase in instrument relevance, that is, when moving right across columns, for both sample sizes. However, the increase is substantially slower than for the LR-test (see first column of Table 2). The latter detects a relevant instrument with 100% probability for both sample sizes if $\rho_1 \geq 0.30$, whereas the F-test does so only in 45% and 75% of the cases, respectively. In other words, the LR-test has more power to reject the null hypothesis when the alternative is true. When comparing both tests, one needs to keep in mind that the LR-test has a priori an advantage as it is based on the correct specification of the time-varying volatility process, whereas the robust F-test accounts for heteroskedasticity of unknown form.[11]

Nevertheless, the suggested decrease in type-II error is useful for practical purposes. It implies that fewer relevant instrument are discarded. The advantage of having an alternative test with more power is also visible when departing from the 5% significance level for the F-test and using the stricter criterion of an F-statistic larger than 10, which is commonly used to shield against weak instrument problems. Table 3 suggests that even when the correlation is $\rho_1 = 0.4$, between 40% and 70% of the relevant instruments are erroneously discarded in samples of 200 and 500, respectively.

[11]The heteroskedasticity-robust F-test is based on a Wald statistic $W \equiv (R\psi - r)'[R\,\mathrm{var}(\psi)R']^{-1}(R\psi - r)$, where ψ is the coefficient of a regression of residuals e_{1t} on a constant and the instrument w_t and $\mathrm{var}(\psi)$ is the variance of ψ. R (r) is a suitable matrix (vector) to restrict ψ to zero.

TABLE 3. Relative rejection frequencies at nominal significance level of 5% of robust F-test for instrument relevance.

		Relevance (β_1, ρ_1)			
Sample Size	Test	(0, 0)	(0.35, 0.15)	(0.72, 0.30)	(1, 0.40)
$T = 200$	Rejection frequency	0.07	0.21	0.45	0.63
	Frequency F > 10	0.01	0.05	0.17	0.30
	Robust F-statistic	1.22	2.57	5.77	8.97
$T = 500$	Rejection frequency	0.05	0.29	0.75	0.91
	Frequency F > 10	0.00	0.06	0.34	0.62
	Robust F-statistic	0.97	3.18	9.18	15.49

Note: The table shows the relative rejection frequencies of robust F-tests for instrument strength at a nominal significance level of 5%, the relative frequencies that $F > 10$, and the average F-statistics, based on 500 replications for each instrument. Endogeneity is assumed to be absent, that is, $\beta_2 = \rho_2 = 0$.

Table 4 displays the evaluation of models A-C using the MSE of the structural impulse responses as accuracy criterion. We normalize the MSE by those of model A and focus on the results based on a sample size of $T = 200$. Online Appendix A.2 shows that the results are robust to changes of the propagation horizon and sample size. For a white-noise instrument ($\rho_1 = \rho_2 = 0$), models A and B yield roughly the same MSE, whereas model C performs extremely poorly. This highlights another attractive feature of our encompassing framework and its usefulness for applied research using external instruments. The result implies that weak instruments are unproblematic for inference if the data contain changes in volatility and if they are used for identification.

For relevant and exogenous instruments, that is, moving south across rows of Table 4, the heteroskedastic proxy-SVAR systematically yields the smallest MSE for all variables and parameterizations. These gains are substantial and increase with instrument relevance. For instruments with a correlation of 0.4 with the monetary shock, the improvement relative to model B is 27% across parameters on average. Model C performs worst in all cases. Given that the variances of the instrument and of the other endogenous variables are time varying in our setup, fitting a standard proxy-SVAR that does not account for heteroskedasticity leads to serious distortions in the estimates of the structural parameters. Overall these results suggest that the explicit modeling of volatility changes when they are a feature of the data and using the information of a valid proxy improves structural inference in SVARs.

This conclusion also holds for slightly endogenous instruments ($\rho_2 = 0.03$) if the proxy is relevant. If $\rho_1 \geq 0.15$, model A consistently yields the smallest MSE. When the endogeneity increases further, the estimation precision of model A deteriorates considerably relative to model B, which ignores the misspecified instrument for identification. Now, model B yields more precise estimates. Its relative MSE are all below one. This finding underscores the importance of being able to test for instrument exogeneity. As before, model C performs worst in all cases. The alternative model C* performs yet worse (Online Appendix A.2), potentially because it requires the estimation of more parameters than the two stage procedure.

TABLE 4. Comparison of MSE of impulse responses to monetary policy shock.

Relevance	Endogeneity (β_2, ρ_2)														
	(0.0, 0.0)			(0.05, 0.03)			(0.17, 0.10)			(0.27, 0.15)			(0.37, 0.20)		
(β_1, ρ_1)	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
(0, 0)															
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	1.02	1.02	1.02	–	–	–	–	–	–	–	–	–	–	–	–
Model C	31.50	24.33	36.21	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.01	1.03	1.02	1.01	1.02	1.03	0.95	0.78	0.92	–	–	–	–	–	–
Model C	16.48	12.73	18.03	16.33	13.06	17.87	15.10	13.64	16.44	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.06	1.20	1.06	1.06	1.18	1.06	0.98	0.90	0.95	0.87	0.54	0.82	0.68	0.31	0.64
Model C	7.91	5.75	7.57	7.85	5.99	7.54	7.73	6.71	7.80	7.24	5.72	7.48	5.89	4.51	6.30
(1.0, 0.40)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.21	1.37	1.23	1.15	1.20	1.14	1.05	1.05	1.04	0.96	0.72	0.92	0.81	0.43	0.75
Model C	5.22	3.62	4.87	5.10	3.58	4.72	5.19	4.46	5.01	5.30	4.45	5.17	4.80	3.68	4.76

Note: The table shows the cumulated MSE of fitted models A–C relative to model A for a propagation horizon up to $h = 25$ and sample size $T = 200$. Each entry is based on 500 replications of each simulation design. Dots $(\cdot)$ denote combinations of values for β_1 and β_2 that produce lower correlations between the instrument w_t and the target structural shock ε_t^r than between the instrument and the nontargeted structural shock ε_t^z. These cases are not taken into account in the analysis.

Summarizing the simulation results, both LR-tests are helpful tools to assess the validity of instruments. Relevant instruments are detected reliably already in small samples at the 5% significance level. Moreover, in our setup the LR-test has more power than the widely used F-test. The detection of endogeneity requires somewhat larger samples and higher correlations between the instrument and the nontargeted shocks. Regarding structural inference, the heteroskedastic proxy-SVAR recovers the true model best, even in cases of slightly endogenous instruments. As endogeneity increases, a standard heteroskedastic SVAR ignoring the instrument for identification performs better and the (heteroskedastic) proxy-SVARs yield seriously distorted estimates. This stresses the importance of having a test for instrument exogeneity. Finally, the heteroskedastic proxy-SVAR yields sharper identification than both alternative models, and the use of heteroskedasticity simplifies the analysis with potentially weak instruments.

3.4 *Simulation results for alternative DGPs*

In this subsection, we modify the set-up of the simulation study along several dimensions to assess the sensitivity of the results and provide practical guidance. First, we modify the changes in volatility and the properties of the instrument to compare the influence of each part on identification. Further, we extend the framework by allowing for two instruments to identify one structural shock. Then we assess the impact of model misspecification. Finally, we take a closer look at weak instruments. In all cases,

we estimate the same models as described in Section 3.2. We present tables for the MSE, focusing on $h = 25$, $T = 200$ and $\rho_1 \leq 0.3$, and summarize the test results verbally. The complete set of test results is in Online Appendix A.2.

First, we model a confounding common shift in the volatility of all shocks by choosing $\Lambda_2 = 2\Lambda_1 + \text{diag}(0.5, 3, 7)$. The top panel of Table 5 shows the normalized MSE. Compared to the baseline simulation, the performance of the encompassing model A improves relative to model B, which is more affected by this shift as it draws only on the heteroskedasticity for identification. The gains of model A over model C decrease somewhat as the advantage of exploiting time-varying volatility declines, but are still sizable.

Alternatively, we model smaller differential changes in the structural shock variances by setting $\Lambda_2 = \text{diag}(0.5, 1, 2)$. Panel 2 of Table 5 shows the results. The gains of model A over model B increase to 76% on average across parameters for exogenous instruments with $\rho_1 = 0.3$. The advantage over model C falls again relative to the baseline simulation, but is still 19% on average for this correlation. For weak and endogenous instruments, model B is best, whereas all models perform roughly similar for strong and endogenous instruments. The latter potentially reflects that less precisely estimated impact effects through heteroskedasticy conflict less with an endogenous instrument.

Next, we generate instruments with censored observations, a common feature of instruments used in empirical applications. We censor the 30%, 60%, or 90% smallest (in absolute value) instrument observations to zero. For 30%, the result hardly changes compared to the baseline. Panel 3 of Table 5 shows the results for a share of 60% zeros. The MSE of models A and B change little. But both heteroskedastic models improve compared to the pure proxy-SVAR, which loses more precision through the censoring. In case of an endogenous instrument, the distortions picked up by model A are less severe such that model B is less advantageous. These patterns extend up to 90% censored instrument observations.[12]

As an extension to the baseline simulations, we consider a framework with two instruments for the monetary shock. Online Appendix A.2 contains the formal details of the model. In the extended framework, the first instrument is generated as described in Section 3.3. As second instrument, we consider two cases. First, we generate a relatively strong and exogenous instrument with $\rho_1 = 0.3$ and $\rho_2 = 0$. We include two instruments in model C accordingly. Panel 4 of Table 5 documents that this improves the performance of models A and C relative to model B in all cases. Even if the first instrument is endogenous, the distortions are less severe if there is a second instrument, which is valid. Alternatively, we construct as second instrument a moderately endogenous one with $\rho_1 = 0.3$ and $\rho_2 = 0.15$ (Panel 5 of Table 5). As expected, the performance gain of model A over B is smaller, given the endogeneity of the second instrument. But the gains of A over C also tend to decline, suggesting that model A is less capable of estimating the larger number of parameters.

The different DGPs affect the performance of the tests for instrument validity typically only mildly. Introducing a confounding common shift or using smaller differential

[12]For the estimation of the MS models based on this simulation design, we shut off the contribution of the instrument to the state probabilities as otherwise we implicitly introduced an additional state via the censored values.

TABLE 5. Comparison of MSE for alternative shock variances and instruments.

Relevance (β_1, ρ_1)	Endogeneity (β_2, ρ_2)														
	(0, 0)			(0.05, 0.03)			(0.17, 0.10)			(0.27, 0.15)			(0.37, 0.20)		
	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
(0, 0)	*Common confounding shift in structural shock variances*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.99	0.93	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model C	17.62	8.46	18.12	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.06	1.14	1.05	0.96	1.00	0.91	0.84	0.59	0.81	–	–	–	–	–	–
Model C	8.86	4.25	8.17	7.91	4.01	6.99	6.99	3.50	6.79	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.27	1.46	1.18	1.28	1.45	1.21	1.10	0.97	0.98	0.91	0.59	0.82	0.74	0.41	0.65
Model C	4.54	2.00	3.74	4.50	2.07	3.77	4.08	2.14	3.47	3.76	2.01	3.47	3.43	2.00	3.22
(0, 0)	*Less distinct changes in structural shock variances*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.98	0.98	0.99	–	–	–	–	–	–	–	–	–	–	–	–
Model C	5.85	3.66	6.38	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.10	1.22	1.05	1.10	1.04	1.05	0.95	0.68	0.93	–	–	–	–	–	–
Model C	2.99	1.75	2.70	2.91	1.61	2.61	2.50	1.58	2.41	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.68	2.04	1.56	1.67	1.96	1.52	1.55	1.45	1.46	1.37	1.01	1.24	1.15	0.71	1.08
Model C	1.54	0.86	1.17	1.54	0.88	1.14	1.45	0.97	1.13	1.32	0.96	1.08	1.13	0.91	1.01
(0, 0)	*60% censored instrument observations*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.93	0.94	0.93	–	–	–	–	–	–	–	–	–	–	–	–
Model C	25.40	16.69	28.60	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.03	0.91	0.96	1.11	0.93	1.05	0.99	0.82	0.96	–	–	–	–	–	–
Model C	12.68	7.56	12.64	13.42	8.15	13.52	11.70	9.20	12.69	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.07	1.07	0.93	0.98	0.96	0.86	0.82	0.78	0.80	0.83	0.60	0.77	0.54	0.37	0.50
Model C	6.02	3.32	5.19	5.52	3.35	4.76	4.79	4.00	4.79	5.21	4.32	5.21	3.66	3.80	3.84
(0, 0)	*Second instrument strong and exogenous ($\rho_1 = 0.3$, $\rho_2 = 0$)*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.74	0.87	0.81	–	–	–	–	–	–	–	–	–	–	–	–
Model C	2.41	3.13	2.58	–	–	–	–	–	–	–	–	–	–	–	–

(*Continues*)

TABLE 5. *Continued.*

Relevance (β_1, ρ_1)	Endogeneity (β_2, ρ_2)														
	(0, 0)			(0.05, 0.03)			(0.17, 0.10)			(0.27, 0.15)			(0.37, 0.20)		
	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	0.74	0.67	0.77	0.87	0.70	0.93	0.81	0.65	0.85	–	–	–	–	–	–
Model C	1.81	2.82	1.88	1.83	2.90	1.92	2.04	3.17	2.19	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.41	1.15	1.42	1.38	1.11	1.43	1.33	1.11	1.39	1.11	0.96	1.14	1.01	0.93	1.00
Model C	1.51	2.01	1.51	1.51	2.06	1.56	1.51	2.52	1.63	1.31	2.64	1.49	1.23	3.08	1.42
(0, 0)	*Second instrument endogenous* ($\rho_1 = 0.3$, $\rho_2 = 0.15$)														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.67	0.54	0.69	–	–	–	–	–	–	–	–	–	–	–	–
Model C	1.72	4.02	1.93	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	0.80	0.70	0.84	0.80	0.63	0.85	0.74	0.58	0.81	–	–	–	–	–	–
Model C	1.42	4.15	1.49	1.53	4.06	1.63	1.52	3.75	1.69	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	0.99	1.36	1.01	0.97	1.30	0.98	0.94	0.99	0.95	0.98	0.82	0.99	1.00	0.61	0.99
Model C	0.92	3.08	0.97	0.94	3.09	1.00	0.99	2.98	1.07	1.08	2.94	1.21	1.20	2.75	1.36

Note: The table shows the cumulated MSE of fitted models A–C relative to model A for a propagation horizon up to $h = 25$ and sample size $T = 200$. Each entry is based on 500 replications of each simulation design. Dots (·) denote combinations of values for β_1 and β_2 that produce lower correlations between the instrument w_t and the target structural shock ε_t^r than between the instrument and the nontargeted structural shock ε_t^z. These cases are not taken into account in the analysis.

changes in shock variances hardly affects the power of the relevance test. The exogeneity test now has a bit more difficulty in detecting endogenous instruments. For the censored instrument, the rejection frequencies are also little affected. For instruments with 60% zeros, the exogeneity test loses power by typically 5 to 10 percentage points (pp) relative to the baseline simulations and the relevance test by less than 5pp. However, for 90% censored instrument observations, both tests lose substantial power (by between 10–40pp). Nevertheless, they still work. In case of a second instrument that is valid, the endogeneity test is unaffected, while the rejection frequencies for the relevance test increase to 98% or more. In other words, the test always detects a relevant instrument when there are one or two relevant instruments. If the second instrument is endogenous, the rejection frequencies for the exogeneity test increase substantially. The lowest fraction is 65% and 98% for the small and large sample, respectively.

To assess the impact of misspecification, we first increase the number of volatility regimes, focusing on the case that the DGP has more states than the fitted MS model.

We introduce $\Lambda_3 = \text{diag}(1, 5, 10)$ and alter the transition matrix to

$$P = \begin{bmatrix} 0.970 & 0.020 & 0.010 \\ 0.025 & 0.950 & 0.025 \\ 0.250 & 0.250 & 0.500 \end{bmatrix}.$$

With this parametrization, the third state resembles a crisis-state with high volatility but low persistence. The first panel of Table 6 shows that the relative performance of the three models is the same as in the baseline simulation design. If the instrument is exogenous, model A typically outperforms models B and C. If the instrument is endogenous, model C is more distorted than before, while there is no clear change in the relative performance of models A and B, which have the same type of misspecification.

Alternatively, we assess the effect of a violation of the assumption of a time-invariant impact effects matrix. We create B_m for $m = 1, 2$ using our baseline specification of B as B_1. To generate B_2, we add a random component to each element of B_1 as follows:

$$B_1 = \begin{bmatrix} 0.69 & 0.61 & 0 \\ -1.10 & 1.49 & 1 \\ -0.75 & 1.49 & 0 \end{bmatrix}, \qquad B_2 = B_1 + \big(\text{vec}(I_3)' \otimes I_3\big)\big(I_3 \otimes \text{vec}\big(\mathcal{N}(0, \sqrt{0.05}I_9)\big)\big),$$

where $\mathcal{N}$ is the normal distribution, I is the identity matrix, $\otimes$ is the Kronecker product, and vec is the vectorization operator. Panel 2 of Table 6 shows that models B and C suffer more than model A from such a misspecification. The latter seems more robust as it draws on two sources of identifying information. Next, we generate time-varying volatility either through a smooth transition in variances DGP using time as transition variable and parameters $c = 0.5T$ and $\gamma = -\sqrt{T}/10$ or through an exogenous break in variances at $0.5T$ (Online Appendix A.4). Panels 3 and 4 of Table 6 show that the MS successfully accounts for different underlying volatility models. The gains of models A and B versus model C increase essentially in all cases as compared to the baseline simulations. Moreover, the performance of model A relative to model B increases in instrument strength as expected even though the advantage of model A over model B declines somewhat. Together, these results suggest that exploiting the heteroskedasticity for identification is easier for the MS model if the DGP has a simpler form of time-varying volatility. This worsens the relative performance of model C and reduces the benefits of A over B.

The LR-tests are robust toward these forms of misspecification. For an underspecified number of states, the rejection frequencies of the exogeneity test fall typically by less than 10pp and those of the relevance test are largely unaffected. In case of a failure of the constancy assumption, the exogeneity test becomes slightly oversized. But the rejection frequencies for endogenous instruments and the relevance test are hardly affected. For the alternative variance models, the rejection frequencies of both tests are essentially unchanged.

Finally, we take a closer look at weak instruments by varying the degree of relevance on a finer grid. Specifically, we simulate exogenous instruments with target correlations of $\rho_1 \in [0, 0.07, 0.1, 0.15, 0.225, 0.3]$ with the monetary shock. 83%–99% of these instruments are rejected in standard proxy-SVARs as being weak according to the typically

TABLE 6. Comparison of MSE when the Markov switching model is misspecified.

	Endogeneity (β_2, ρ_2)														
Relevance	(0, 0)			(0.05, 0.03)			(0.17, 0.10)			(0.27, 0.15)			(0.37, 0.20)		
(β_1, ρ_1)	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
(0, 0)	*DGP is Markov switching model with 3 states*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	1.01	1.02	1.02	–	–	–	–	–	–	–	–	–	–	–	–
Model C	29.58	19.18	32.32	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.06	1.14	1.11	1.07	1.10	1.11	1.40	1.03	1.42	–	–	–	–	–	–
Model C	18.91	12.66	19.99	19.58	12.94	20.81	19.07	14.77	20.91	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	0.85	1.18	0.83	1.13	1.24	1.19	1.03	1.05	1.07	0.89	0.65	0.89	1.08	0.50	1.04
Model C	7.11	4.90	6.56	9.62	6.14	9.59	7.31	7.07	7.54	6.69	6.70	7.09	8.11	6.47	8.67
(0, 0)	*DGP has small shift in impact matrix across regimes*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	0.99	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model C	31.89	21.63	37.21	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.04	1.08	1.07	1.06	1.11	1.06	1.00	0.83	0.97	–	–	–	–	–	–
Model C	18.40	11.70	21.26	19.02	12.87	21.27	16.73	13.27	19.26	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.16	1.33	1.16	1.15	1.25	1.14	1.04	0.93	1.02	0.98	0.60	0.94	0.79	0.36	0.74
Model C	8.19	4.55	7.98	8.12	4.89	7.93	7.94	6.10	8.17	8.16	6.02	8.68	6.40	4.88	6.91
(0, 0)	*DGP with smooth transition in variances*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	1.00	1.00	1.01	–	–	–	–	–	–	–	–	–	–	–	–
Model C	33.55	30.85	39.80	–	–	–	–	–	–	–	–	–	–	–	–
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.03	1.02	1.03	1.03	1.01	1.02	1.01	0.93	0.99	–	–	–	–	–	–
Model C	23.55	20.99	27.67	22.94	21.46	27.09	21.59	24.93	26.21	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.06	1.04	1.06	1.05	1.03	1.05	0.99	0.84	0.97	0.91	0.63	0.87	0.74	0.37	0.68
Model C	11.14	8.82	11.97	11.07	9.45	12.03	11.03	11.28	12.44	10.85	11.89	12.67	9.26	9.39	10.76
(0, 0)	*DGP has exogenous break in variances*														
Model A	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model B	1.00	1.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
Model C	38.57	34.64	45.57	–	–	–	–	–	–	–	–	–	–	–	–

(*Continues*)

TABLE 6. *Continued.*

Relevance	Endogeneity (β_2, ρ_2)														
	(0, 0)			(0.05, 0.03)			(0.17, 0.10)			(0.27, 0.15)			(0.37, 0.20)		
(β_1, ρ_1)	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
(0.35, 0.15)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	–	–	–	–	–	–
Model B	1.02	1.00	1.02	1.02	0.99	1.02	1.00	0.91	0.99	–	–	–	–	–	–
Model C	26.69	22.93	31.27	25.96	23.57	30.58	24.59	27.55	29.81	–	–	–	–	–	–
(0.72, 0.30)															
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.06	1.03	1.06	1.05	1.01	1.05	1.01	0.83	0.99	0.91	0.62	0.87	0.76	0.37	0.69
Model C	12.66	10.12	13.70	12.60	10.55	13.75	12.92	12.53	14.47	12.43	13.23	14.39	10.73	10.46	12.49

Note: The table shows the cumulated MSE of fitted models A–C relative to model A for a propagation horizon up to $h = 25$ and sample size $T = 200$. Each entry is based on 500 replications of each simulation design. Dots (·) denote combinations of values for β_1 and β_2 that produce lower correlations between the instrument w_t and the target structural shock ε_t^r than between the instrument and the nontargeted structural shock ε_t^z. These cases are not taken into account in the analysis.

used robust F-statistic (Table 3). In contrast, the heteroskedastic proxy-SVAR allows using all of them. Table 7 shows that the MSE of model A are at least as small as those for model B, and always smaller than the ones for model C. The advantage of model A over B increases in instrument relevance, whereas the improvement over model C decreases. Taken together, these results suggest that one can include a weak instrument into a heteroskedastic SVAR to label the main shock of interest without blurring inference, and which in many cases may sharpen inference.

4. MONETARY POLICY ANALYSIS IN A HETEROSKEDASTIC PROXY-SVAR

We use our framework to provide new—and in light of the Monte Carlo evidence sharper and more reliable—estimates of the impact of monetary policy on the macroeconomy. Our baseline model consists of four endogenous variables and an instrument for monetary policy shocks in the vector $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$. We use the federal funds rate as the monetary policy indicator ff_t, the log of industrial production as a measure of real economic activity ip_t, the log of the personal consumption expenditure core price index pce_t as a measure of the Federal Reserve price target variable, and the log of a price index of raw materials prm_t to deal with anticipation of future inflation and the

TABLE 7. Comparison of MSE of for weak instruments.

Relevance	(0, 0)			(0.16, 0.07)			(0.23, 0.1)			(0.35, 0.15)			(0.53, 0.225)			(0.72, 0.30)		
(β_1, ρ_1)	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}	θ_{11}	θ_{21}	θ_{31}
Model A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model B	1.02	1.02	1.02	1.02	1.03	1.03	1.00	1.02	1.00	1.01	1.03	1.02	1.02	1.06	1.03	1.06	1.20	1.06
Model C	31.50	24.33	36.21	26.20	19.83	29.62	22.15	16.59	23.83	16.48	12.73	18.03	11.11	8.29	11.37	7.91	5.75	7.57

Note: The table shows the cumulated MSE of fitted models A–C relative to model A for a propagation horizon up to $h = 25$ and sample size $T = 200$. Each entry is based on 500 replications.

TABLE 8. Model selection.

Reduced Form Models	$\log(L_T)$	SC	AIC	HQ
Markov switching 3 states	2637.660	−4883.320	−4587.544	−4142.295
Smooth transition in variances (time)	2578.269	−4782.538	−4500.343	−4075.539
Markov switching 2 states	2577.234	−4780.468	−4498.272	−4073.469
GARCH residuals	2498.744	−4617.487	−4330.765	−3899.146
Smooth transition in variances (IP)	2441.076	−4508.152	−4225.957	−3801.153
Exogenous breakpoint	2427.803	−4483.607	−4202.921	−3780.388
White noise residuals	2290.697	−4241.394	−3984.853	−3598.667

Note: L_T denotes the likelihood function evaluated at the optimum, $AIC = -2\log(L_T) + 2f$, $SC = -2\log(L_T) + \log(T)f$ and $HQ = -2\log(L_T) + 2f \times \log(\log(T))$, where f is the number of free parameters and T the number of observations.

associated price puzzle, following Christiano, Eichenbaum, and Evans (1999).[13] We take the narrative-based measure of unexpected changes in the intended fed funds rate of Romer and Romer (2004) as an instrument for the latent monetary policy shocks, rr_t.[14] This proxy starts in 1969M1 and has the longest sample, while instruments constructed with high-frequency data usually start only in the 1990s.

We estimate the VAR on monthly frequency data within the sample 1980M1 to 2007M6. The start is dictated by the availability of the raw materials price data, while the end is chosen such as to ensure that our sample is not affected by the zero lower bound or by unconventional monetary policy. We use six lags to account for the persistence in the data. In Online Appendix A.4, we conduct and extensive robustness analysis and show that our results hold when changing the number of lags and states, the sample period, the monetary policy indicator, and an alternative volatility model.

4.1 *Model specification*

An important choice in our framework is the volatility model. Its functional form affects the likelihood, estimators, and tests. Therefore, we perform an extensive model comparison. As candidates, we describe heteroskedasticity through smooth transition in variances using either a 12-month trailing moving average of industrial production or time as the transition variable, an exogenous break point iterating over all potential break points, a multivariate GARCH process, as well as MS models with $M = 2$ and $M = 3$ states, respectively.[15] Online Appendix A.4 describes all models formally.

Table 8 shows the log-likelihood for a linear model assuming white noise residuals and the five alternative volatility models. In addition, we report information criteria because they work well for judging the performance of MS models (Psaradakis and Spagnolo (2006)), whereas standard tests are problematic for this purpose as some parameters might not be identified under the null hypothesis of a smaller number of states

[13]The series are *INDPRO*, *FEDFUNDS*, *PCEPILFE*, and *PINDUINDEXM* downloaded from the (AL)FRED database of the Federal Reserve Bank of St. Louis.

[14]We use the updated version of the original Romer and Romer (2004) constructed by Wieland and Yang (2016) downloaded from Wieland's webpage.

[15]We also estimate smooth transition in variances models using either 6- or 24-month moving averages of industrial production. Both perform worse than the 12-month version so we do not report them.

than under the alternative (Hansen (1992)). Specifically, the AIC chooses successfully between alternative volatility models Lütkepohl and Schlaak (2018). The models are ordered descending according to their log-likelihood values. This ranking coincides with that implied by the information criteria.

The table conveys two results. First, the linear model is dominated by all models that allow for changes in volatility. This strongly supports the assumption of heteroskedasticity. In this case, using any time-varying volatility model estimates structural impulse responses more precisely than a linear model (Lütkepohl and Schlaak (2018)). Second, the MS models tend to be preferred over the other heteroskedastic models. We opt for the MS(2) model. It yields more stable and precise estimates given that the third state in the MS(3) model contains only few observations. Moreover, relative to the smooth transition in time model, MS models are usually the best choice even in cases where the volatility specification does not coincide with the data generating process (Lütkepohl and Schlaak (2018)). Modeling changes in volatility through a latent variable gives full voice to the data, reducing the risk of misspecification of the transition variables, functions, or break points. In Online Appendix A.4, we show that the results are robust to using the MS(3) or the smooth transition model.

At the same time, MS models have several limitations due to their computational complexity. First, they require sufficiently many observations (in each regime) to estimate the parameters reliably. While there is no formal analysis, the existing frequentist applications suggests that the minimum sample length is 100 for a bivariate model (Podstawski and Velinov (2018)). In our simulations, we obtain reliable results for a four-variable model with 200 observations. Second, the MS models cannot accommodate a large number of variables. For our empirical application, we need a high performance computing server in order to check the optima of the likelihood function of a sufficiently high number of starting values to assure that a global optimum is found. Hence, five endogenous variables is probably an upper bound for frequentist applications in practice. Bayesian MS models may accommodate up to six variables (Lütkepohl and Woźniak (2020)). Similarly, the number of lags in our application (6) rather constitutes a maximum, while other studies typically use fewer lags (Podstawski and Velinov (2018), Herwartz and Lütkepohl (2014), Lütkepohl and Woźniak (2020), Lanne, Lütkepohl, and Maciejowska (2010), Lütkepohl and Schlaak (2018)). Summarizing, the strong nonlinearity of the MS model may limit its applicability if the data or research question require many endogenous variables to avoid nonfundamentalness, or many observations to have sufficient data points in each regime. Solutions to these bottlenecks are high performance computing servers, Bayesian analysis, or higher frequency data.

4.2 *Volatility regimes and identification*

Table 9 reports the estimated state-dependent reduced form covariance matrices. They indicate whether the model detects switches in volatility, which are central for identification and testing. They also help interpret the endogenously and agnostically identified regimes. The variances all increase in state 2, by factors of 2, 53, 5, 3, and 18 across rows. The strong increase in the volatility of the interest rate residual by 53 is further evidence

TABLE 9. Estimated state covariance matrices ($\times 10^3$) of reduced form model (8) with $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$.

State 1: $\tilde{\Sigma}_1$

$$\begin{bmatrix} 0.443 & & & & \\ 0.090 & 1.562 & & & \\ 0.003 & 0.005 & 0.017 & & \\ 0.404 & 0.295 & 0.024 & 9.526 & \\ -0.375 & 5.518 & 0.026 & -0.604 & 157.933 \end{bmatrix}$$

State 2: $\tilde{\Sigma}_2$

$$\begin{bmatrix} 0.821 & & & & \\ 3.389 & 82.749 & & & \\ -0.005 & -0.262 & 0.089 & & \\ -0.161 & -1.311 & -0.161 & 32.801 & \\ 24.978 & 277.004 & -0.709 & 7.996 & 2845.134 \end{bmatrix}$$

that the sample is characterized by changes in monetary policy volatility. Moreover, the MS model seems able to detect and separate these changes.

The table also shows that the covariances increase (in absolute value) in state 2, and often by larger factors than the variances. These changes in the covariances illustrate the idea behind identification through heteroskedasticity. In a period where interest rates are highly volatile, we learn more about the relation between the federal funds rate, economic activity, and prices as the covariance between them temporarily increases. Monetary policy shocks are then more likely to occur and can be used as a "probabilistic instrument" (Rigobon (2003)) to trace out the response of production and prices.

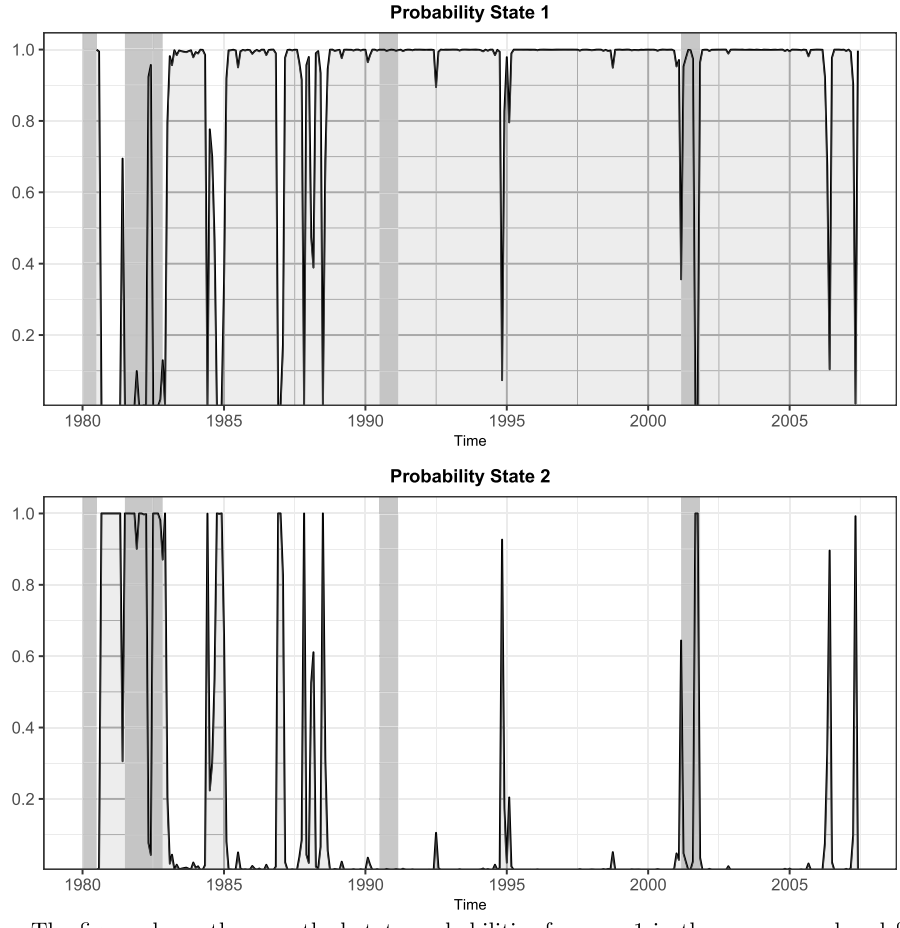
To test the validity of the external instrument, we use statistical identification. The latter requires significant and differential changes in the volatility of the structural innovations μ_t. Table 10 shows the estimated variances of the structural model (10) with unrestricted β in state 2, which are contained in Λ_2. Given the restrictions on D and that the instrument is ordered last in z_t, λ_{52} captures the change in the variance of the noise in the measurement of the instrument. As the ordering of the remaining $\lambda_{\cdot 2}$s is arbitrary, we simply order them from largest to smallest. All estimates are larger than one, implying that all structural shocks are more volatile in state 2. Thus, we label state 2 the high volatility state. Identification requires that the variance shifts are all distinct from each other. The heterogeneity of the elements of Λ_2 points toward statistical identification. The confidence intervals constructed using one standard deviation around the point estimates do not overlap. This suggests that the decomposition in (11) is locally unique and can be used to test the validity of the instrument.

To develop an economic notion about the statistically identified regimes, Figure 1 shows the smoothed state probabilities. The upper panel corresponds to state 1 and the

TABLE 10. Estimates and standard errors of relative variances.

Parameter	Estimate	Standard erro r
λ_{12}	14.56	3.26
λ_{22}	5.56	1.27
λ_{32}	3.49	0.78
λ_{42}	1.23	0.28
λ_{52}	53.59	12.18

Note: The standard errors are obtained from the inverse of the negative Hessian evaluated at the optimum of the structural model (10) with $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$.



Notes: The figure shows the smoothed state probabilities for $m = 1$ in the upper panel and for $m = 2$ in the lower panel of model (10) with $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$. The shaded vertical bars mark recession periods defined by the NBER.

FIGURE 1. Smoothed state probabilities.

lower panel to state 2. State 1 dominates the sample. The model detects a long spell of low volatility during a period that is often referred to as the "Great Moderation" in the 1990s and 2000s with stable growth and inflation under the Federal Reserve Chairman, Alan Greenspan. The high volatility regime appears more often during the first part of the sample. Many of the spikes are associated with specific events in the economic history of the U.S. In particular, a longer-lasting switch to state 2 coincides with the chairmanship of Paul Volcker during the first half of the 1980s. In the second part of the sample, there are peaks around the burst of the dot-com bubble in 2001, the 9/11 attacks, and the subsequent recession. Overall, this narrative, while only suggestive, indicates that the endogenously determined volatility regimes capture relevant developments in the U.S. economy and in the conduct of monetary policy.

In model (10), we assume that both the autoregressive parameters $\Gamma(L)$ and the structural impact matrix D are state-invariant. The first assumption ensures compu-

TABLE 11. Instrument validity.

	Exogeneity	Relevance
LR statistic	0.80	44.85
Degrees of freedom	3	1
p-value	0.85	0.00

Note: The table shows the LR statistic, the p-value, and the number of restrictions for the tests of instrument exogeneity ($H_0 : \beta_2 = \cdots = \beta_K = 0, H_1 : \beta$ unrestricted) and instrument relevance ($H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$). The instrument is the narrative-based measure of monetary surprises of Romer and Romer (2004).

tational tractability. The latter is a crucial ingredient for identification in our setup and standard in the literature on identification via heteroskedasticity.[16] These assumptions also imply that the model attributes time-varying reduced form volatility entirely to the changes in structural shock variances and not to changes in the policy rule of the estimated monetary SVAR. The evidence in Owyang and Ramey (2004), Primiceri (2005), Sims and Zha (2006), and Amir-Ahmadi, Matthes, and Wang (2016) supports this assumption. These papers examine the drivers of volatility changes in the U.S. over time. While they find some evidence for regime switches in the policy rule, they conclude that these changes explain only a small part of U.S. business cycles and that changes in shock variances explain most of the time-varying volatility. Other authors find little or no evidence of changes in the monetary policy coefficients (Bernanke and Mihov (1998), Leeper and Zha (2003). Recently, Antolín-Díaz and Rubio-Ramírez (2018)) follow these views by assuming no variation in the policy rule in their monetary SVAR.

Our framework allows testing the assumption of a constant impact matrix implicitly. To assess whether our assumption of zeros on the last column of D is in line with the data (9), we test the null hypothesis $H_0 : d_{15} = d_{25} = d_{35} = d_{45} = 0$ against the alternative hypothesis $H_1 : \exists j \in \{1, \ldots, 4\}$s.t.$d_{j5} \neq 0$. The associated LR-test has a $\chi^2(4)$ distribution. The LR-statistic is 6.582 with a p-value of 0.160, not rejecting the zero restrictions. The test result also implies that there is no evidence against the constancy of B and β across states because the data could speak up against this assumption in an overidentified system.

4.3 *Instrument validity*

We now use the significant and distinct changes in the variances of the structural innovations to test the validity of the instrument. First, we test for exogeneity. We test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ against $H_1 : \beta$ unrestricted. Thereby, we order the monetary shock first and test whether the instrument is exogenous to the nonmonetary policy shocks. Table 11 shows that the data do not reject the assumption of exogeneity. The LR-statistic is small and the p-value far above conventional significance levels. Thus, the instrument can be considered as exogenous in our model.

Second, we test for relevance. We test the null hypothesis that the instrument is unrelated to all structural shocks, $H_0 : \beta = 0$, against the alternative that it is significantly

[16]See Sentana and Fiorentini (2001), Rigobon (2003), Rigobon and Sack (2003), Normandin and Phaneuf (2004), Herwartz and Lütkepohl (2014), and Nakamura and Steinsson (2018).

related to at least one structural shock. If the null is rejected, this will be the monetary policy shock given the first stage result and that the instrument is constructed to have a high correlation with the monetary shock and a low (zero) correlation with the other shocks. The test indicates that the instrument is highly relevant. The null is rejected at the 1% significance level.[17] We conclude that the monetary surprises of Romer and Romer (2004) are a valid instrument for monetary policy shocks in our SVAR.

These results, as any tests, are model-specific. Nevertheless, they suggest that the measure of Romer and Romer (2004) successfully addresses endogeneity concerns raised by Leeper (1997) about their earlier narrative measures of presumably exogenous policy changes (Romer and Romer (1989)). Moreover, the results indicate that the measure is a valid instrument for monetary policy shocks. This supports the results of many papers employing it as such in time-series models (Stock and Watson (2012), Tenreyro and Thwaites (2016), Ramey (2016), Rey (2016)).

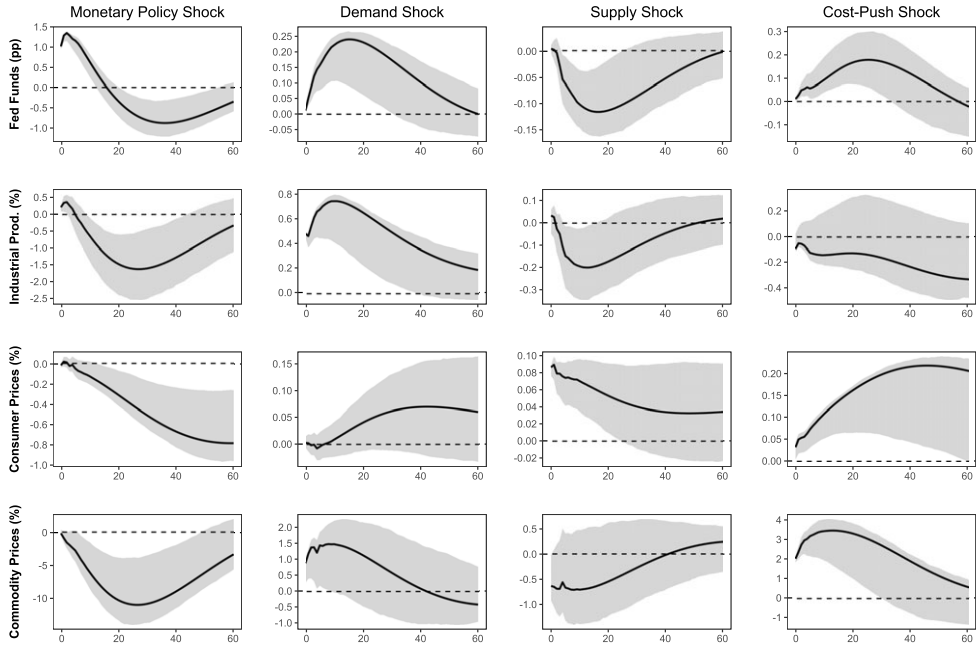
4.4 *Dynamic effects of monetary policy shocks*

We now estimate the dynamic effects of monetary shocks. Based on the testing sequence, we leave β_1 unrestricted and set $\beta_2 = \beta_3 = \beta_4 = 0$. This implies that the estimation combines the information in the instrument and the time-varying second moments for identification of a heteroskedastic proxy-SVAR.

Figure 2 shows the impulse responses to all four structural shocks in columns on the endogenous variables in rows. When interpreting the results, the inclusion of the proxy is a key advantage over traditional identification through heteroskedasticity, where a main challenge is the economic interpretation of the statistically identified shocks (Herwartz and Lütkepohl (2014)). The restrictions on β pin-down the monetary shock in the first column of D. A 100 basis points surprise monetary contraction leads to an significantly elevated federal funds rate for a year and a half, according to the 95% confidence intervals. Economic activity declines half a year after the occurrence of the shock. The response bottoms after 2 years and is highly statistically significant. The trough is -1.6%. Thereafter, industrial production gradually returns to its trend, but is still depressed at the end of the propagation horizon. Consumer prices fall steadily after the shock, but more sluggishly and persistently than activity. The trough is -0.8% after 5 years. With the weakening economy, the policy rate falls below trend after 2 years, reflecting an endogenous response of monetary policy. Finally, raw material prices fall significantly.

Qualitatively, the results are in line with a long literature that documents contractionary effects of unexpected increases in the federal funds rate on output and prices (Christiano, Eichenbaum, and Evans (1999), Gertler and Karadi (2015), Caldara and Herbst (2019)). Importantly, there is no price puzzle. Quantitatively, the estimated impacts are "medium" in the terminology of Coibion (2012), that is, they are roughly inbetween the small effects of monetary policy shocks documented by Christiano, Eichenbaum, and Evans (1999) and the large effects presented by Romer and Romer (2004). Compared to the latter paper, our estimates for industrial production are only about

[17]The alternative relevance test yields the same conclusion. The LR-statistic is 45.65 with 4 degrees of freedom. The corresponding p-value is virtually zero, clearly rejecting the null hypothesis of irrelevance.



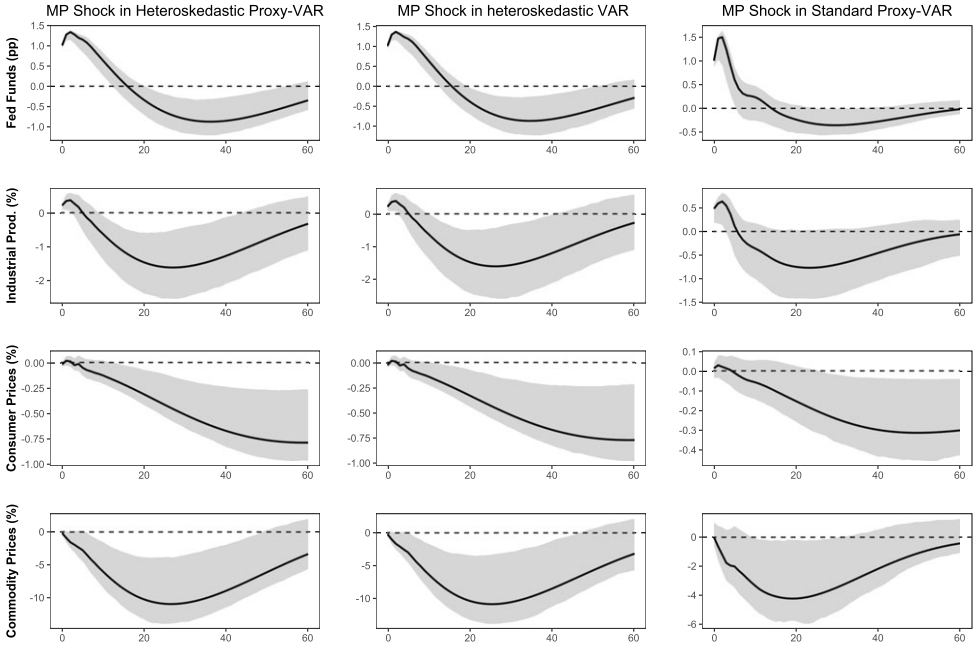
Notes: The figure shows the impulse responses to one standard deviation shocks in state $m = 1$ of the heteroskedastic proxy-SVAR(6) model with $M = 2$ states for $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$. The sample is 1980M1–2007M6 and the instrument for monetary policy shocks is the narrative-based measure of Romer and Romer (2004). The shaded bands denote 95% pointwise confidence intervals based on 1,000 bootstrap replications.

FIGURE 2. Impulse responses for heteroskedastic proxy-SVAR model.

half of the size. We attribute this to the more recent sample, as the impact of monetary surprises is typically estimated to be at least partly smaller after the 1980s (Ramey (2016)).

The other structural shocks are identified using the changes in volatility and yet need to be labeled (if they are of interest to the researcher). Our framework also simplifies this task compared to traditional identification through heteroskedasticity. The inclusion of a valid proxy separates the shock of interest from the remaining shocks. Therefore, the latter are easier to label. This is reflected in relatively clear sign patterns of the impulse responses for the other shocks and in the forecast error variance decomposition (Online Appendix A.3), which both suggest a clear economic labeling of the shocks.

The second shock leads to an increase in economic activity, consumer prices, and the federal funds rate. These signs are consistent with a demand shock that induces tighter policy. The third shock implies a significant decline in real activity and a concurrent increase in consumer prices. Thus, we label it a supply shock. Finally, the fourth shock explains roughly 80% of the error variance in raw material prices. It leads to an immediate and significant jump in this price index that feeds into consumer prices with a delay of several months. This is associated with a concurrent increase the federal funds



Notes: The figure shows the impulse responses to a monetary policy shock of 1pp in a heteroskedastic proxy-SVAR (left column), a heteroscedastic SVAR (middle column), and a proxy-SVAR (right column) on the endogenous variables in rows. The model contains $z_t = [ff_t, ip_t, pce_t, prm_t, rr_t]'$, the sample is 1980M1–2007M6 and the instrument for latent monetary shocks is the narrative-based measure of Romer and Romer (2004). The shaded bands denote 95% pointwise confidence intervals based on 1,000 bootstrap replications.

FIGURE 3. Comparison of models.

rate. Industrial production falls given higher input costs and tighter policy. These patterns reflect those induced by a cost-push shock.

The simulation study suggests that another advantage of modeling heteroskedasticity is sharper identification of the structural model, and hence of the effects of monetary policy. To see this in the application, Figure 3 compares the impulse responses to a monetary policy shock from the heteroskedastic proxy-SVAR (left column) to those from a standard heteroskedastic SVAR with $\beta = 0$ (middle column) and from a standard proxy-SVAR neglecting time-varying volatility (right column). The shock is scaled to 100 basis points for comparison. Qualitatively, all models yield the same conclusions. Production and prices decline.

Quantitatively, however, and in terms of statistical significance, there are notable differences. In the heteroskedastic SVAR models, the monetary shock is about twice as persistent. The federal funds rate remains significantly above trend for about 18 months, whereas in the standard proxy-SVAR it is indistinguishable from zero after roughly three quarters. This stronger and longer-lasting monetary contraction leads to a quicker, stronger, and more persistent drop in industrial production. In the heteroskedastic SVAR models, output falls significantly below trend after a year and remains depressed for

more than 3 years. The trough is at −1.6%. In contrast, in the proxy-SVAR, the decline in economic activity is largely insignificant and the trough is only −0.8%. Similarly, the effect of the monetary shock on prices is quicker, about twice as strong, and more statistically significant in the heteroskedastic models.

The comparison suggests that the heteroskedastic proxy-SVAR attributes a larger role to monetary shocks in business cycles than the standard proxy-SVAR. To see whether this is the case, we compute forecast error variance decompositions. Indeed, the heteroskedastic proxy-SVAR implies that monetary shocks account for 68%–76% of the long-run variation (at horizon 60) of industrial production and both price measures in the high volatility regime, and for 7%–15% in the low volatility regime (Online Appendix A.3). The standard proxy-SVAR implies that monetary shocks explain 9%–18% of the variance of these variables. When interpreting the numbers of the heteroskedastic proxy-SVAR, one needs to keep in mind, however, that the high volatility regime accounts only for one-fifth of the observations.

Finally, the similarity between the impulse responses for the first two models suggest that heteroskedasticity provides strong identifying information in the sample. In other words, the inclusion of the instrument changes the responses only marginally. This observation, in turn, has two implications. First, in this application, the main advantage of including the instrument is to pin-down the shock of interest. Second, heteroskedasticity alone can in principle be used to identify monetary policy shocks in U.S. post WWII samples. The statistically identified shocks are consistent with those that also use an instrument for identification. This supports the results from a long series of papers that use time-varying volatility to identify monetary policy shocks (Normandin and Phaneuf (2004), Lanne and Lütkepohl (2008), Lütkepohl and Woźniak (2020)).

4.5 *Testing alternative proxies for monetary shocks*

Finally, we test and compare further measures of monetary surprises proposed in the literature. We consider the identified monetary shocks from the SVAR of Bernanke, Boivin, and Eliasz (2005, BBE05) and monetary surprises identified using high(er) frequency data. For the latter, we employ measures derived from changes in federal funds futures data around policy announcements using a daily window (Barakchian and Crowe (2013, BC13)), a 30-minute window (Gertler and Karadi (2015, GK15)), and a 30-minute window including further cleaning of the surprises by regression on a range of control variables (Miranda-Agrippino and Ricco (2018, MR18)). Finally, we compute the first principal component of all instruments, which accounts for about 44% of their variation, to see whether combining the information from multiple instruments generates an advantage. We consider the potential instruments one at a time. To facilitate a clean comparison, we use the baseline sample period for the evaluation of all proxies although they are available for slightly different periods. We fill the missing observations with zeros.

Table 12 shows the test results. The model-based measure is a valid instrument. The p-values are 0.68 for exogeneity and essentially zero for relevance. The picture is more mixed for the instruments based on high-frequency data. In particular, for the instrument based on daily data, there is some indication of endogeneity. The p-value of 0.127

TABLE 12. Testing alternative proxies.

Instrument	Exogeneity p-Value	Relevance p-Value
Model-based (BBE05)	0.680	0.000
Daily (BC13)	0.127	0.593
30-minute (GK15)	0.608	0.056
30-minute cleansed (MR18)	0.584	0.036
Factor	0.673	0.000

Note: The table shows the p-values of LR-tests for the exogeneity and relevance of different instruments w_t in the model with $z_t = [ff_t, ip_t, pce_t, prm_t, w_t]'$, testing them one at a time. The sample period is 1980M1–2007M6. The model-based measure is of Bernanke, Boivin, and Eliasz (2005), and measures based on high(er) frequency data are taken from Barakchian and Crowe (2013), Gertler and Karadi (2015), and Miranda-Agrippino and Ricco (2018), respectively, and the first principal component of these instruments.

is only marginally above the 10% level that the simulations suggest using. Moreover, the hypothesis of irrelevance cannot be rejected. The p-value of 0.593 is quite large. The remaining three instruments are all exogenous, according to the high p-values of close or above 0.5. But the 30-minute instrument and the cleansed version thereof are not particularly strong, with p-values of 0.056 and 0.036, respectively. The relevance tests for the higher-frequency instruments need to be treated with some caution, however, due to the smaller number of nonzero instrument observations.

We compare our findings based on the LR-tests to three alternative procedures that are used to evaluate the quality of instruments. First, we use Granger-causality (GC) tests to see whether lags of the endogenous variables predict the instruments, thereby following procedure based on joint correlations (Miranda-Agrippino and Ricco (2018)). Projecting one instrument at a time onto the autoregressive part of the VAR, we test the null hypothesis that the lags of each endogenous variable are jointly equal to zero. Table 13 shows robust p-values of the corresponding F-tests and some summary statistics. The R^2s are modest at around 10%-20%, suggesting little explanatory power of all 30 lags for the instruments. Moreover, there is generally little evidence of Granger causality. The baseline instrument (RR04) and the daily surprises (BC13) are not Granger caused by any of the variables. The model-based instrument (BBE05), the cleansed 30-minute surprises (MR18), and the factor are Granger caused only by one variable at the 10% significance level. An exception are the 30-minute surprises (GK15), which are Granger caused by three variables at the 5% significance level. Accordingly, the R^2 is 0.38.

Second, we use the overidentification test proposed by Cesa-Bianchi, Thwaites, and Vicondoa (2016), which is based on the availability of two instruments. For each reduced form residual $e_{2t}, \ldots, e_{4t}$, we compute Hansen's J-statistics testing the joint null hypothesis that two instruments are valid. We keep the baseline instrument of Romer and Romer (2004) fixed across tests and add one alternative instrument at a time. Table 14 reports the p-values of the (HAC-consistent) J-tests. The results add to the evidence based on the LR-tests and GC-tests. The p-values are all large for the model-based measure and the factor. For the daily and 30-minute proxy, the test rejects instrument validity at the 5% and 10% level, respectively, when using the residual of the equation of consumption expenditure prices. The cleansed 30-minute instrument is a borderline case.

TABLE 13. p-values of Granger-causality tests.

Dependent Variable	RR04	BBE05	BC13	GK15	MR18	Factor
p-value GC-test						
Instrument	0.69	0.47	0.36	0.04	0.04	0.58
Federal funds rate	0.12	0.83	0.67	0.00	0.48	0.83
ln(Industrial production)	0.87	0.99	0.51	0.29	0.29	0.98
ln(PCE core)	0.60	0.13	0.58	0.03	0.19	0.37
ln(Material prices)	0.47	0.02	0.89	0.46	0.41	0.09
Observations	384	209	235	263	233	203
Parameters	31	31	31	31	31	31
R^2	0.20	0.12	0.13	0.38	0.23	0.10

Note: The table shows p-values of robust F-statistics using HAC standard errors testing the null hypothesis that the coefficients on six lags of each variable in the VAR with $z_t = [ff_t, ip_t, pce_t, prm_t, w_t]'$ are jointly equal to zero. The instruments w_t are included one at a time into the VAR and are RR04—Romer and Romer (2004), BBE05—Bernanke, Boivin, and Eliasz (2005), BC13—Barakchian and Crowe (2013), GK15—Gertler and Karadi (2015), MR18—Miranda-Agrippino and Ricco (2018), and the first factor of these.

Third, we regress the instruments on different types of structural shocks available from the literature to see whether the instruments are potentially endogenous to these shocks. We use all shocks and variables considered by Stock and Watson (2012) except for monetary policy shocks and a few measures with insufficient observations. We consider oil shocks (Hamilton (2003), Kilian (2008), Ramey and Vine (2011)), uncertainty shocks (Baker, Bloom, and Davis (2016)), fiscal policy shocks (Romer and Romer (2010), Fisher and Peters (2010), Ramey (2011), productivity shocks Basu, Fernald, and Kimball (2006), Smets and Wouters (2007)), and financial shocks (Gilchrist and Zakrajsek (2012)). Half of these series are available only for the quarterly frequency so we sum the instruments within quarters. We regress one proxy on all potential shocks for a given frequency at a time.

Table 15 shows the regression results with HAC standard errors. The results are similar to those based on the LR-tests. The R^2s are typically low, indicating little relevance of the shocks. Moreover, out of the 36 coefficients, only 8 are statistically significant, thereof three only at the 10% level. An exception are the higher frequency measures that are not cleansed. They seem to systematically relate to oil and uncertainty shocks, consistent with the indication of endogeneity by the LR-test and that they are Granger caused by

TABLE 14. p-values of Hansen's overidentification J-test.

Residual	BBE05	BC13	GK15	MR18	Factor
ln(Industrial production)	0.85	0.64	0.40	0.96	0.44
ln(PCE core)	0.42	0.07	0.03	0.10	0.45
ln(Material prices)	0.89	0.86	0.61	0.39	0.75

Note: The table shows p-values of Hansen's HAC-consistent J-statistics testing the joint null hypothesis that the instruments are valid. The instruments are tested one at a time against the baseline instrument of Romer and Romer (2004). The alternative instruments are BBE05—Bernanke, Boivin, and Eliasz (2005), BC13—Barakchian and Crowe (2013), GK15—Gertler and Karadi (2015), MR18—Miranda-Agrippino and Ricco (2018), and the first factor of these.

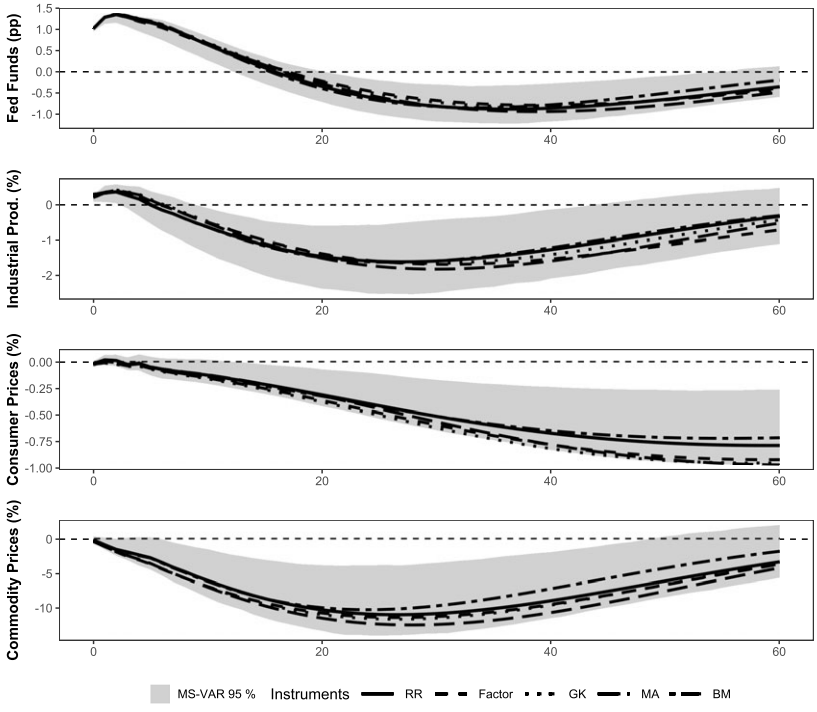
TABLE 15. Regression of proxies on other shocks.

Dependent Variable	RR04	BBE05	BC13	GK15	MR18	Factor
(a) Monthly data						
Ramey and Vine (2011)	1.39	14.98	30.87	3.34	1.25	1.90
	(6.59)	(49.39)	(27.36)	(1.29)	(2.05)	(3.87)
Kilian (2008)	−0.04	−0.01	−0.29	0.01	0.00	−0.01
	(0.06)	(0.46)	(0.30)	(0.02)	(0.02)	(0.03)
Hamilton (2003)	0.19	0.62	2.55	−0.01	0.09	0.17
	(0.33)	(2.70)	(2.15)	(0.07)	(0.14)	(0.18)
TED spread	−3.87	−34.41	−2.19	−1.11	0.04	0.33
	(3.41)	(27.87)	(15.02)	(0.68)	(1.23)	(2.20)
Baker, Bloom, and Davis (2016)	0.02	−0.36	−0.73	−0.03	−0.01	0.00
	(0.05)	(0.34)	(0.27)	(0.01)	(0.02)	(0.02)
AR(2) residual of VIX	0.26	0.42	3.53	0.28	0.04	−0.16
	(0.32)	(2.36)	(1.70)	(0.08)	(0.12)	(0.20)
Observations	321	215	241	260	239	209
R^2	0.09	0.06	0.10	0.11	0.06	0.06
(b) Quarterly data						
Ramey (2011)	2.13	20.68	2.41	3.16	2.15	0.89
	(9.43)	(65.46)	(12.52)	(5.44)	(3.72)	(2.73)
Fisher and Peters (2010)	2.31	2.54	−1.28	0.19	0.08	−0.04
	(1.33)	(3.58)	(2.83)	(0.32)	(0.24)	(0.14)
Romer and Romer (2010)	−0.18	1.01	0.11	−0.12	0.00	−0.03
	(0.29)	(1.50)	(0.55)	(0.14)	(0.08)	(0.14)
Smets and Wouters (2007)	−0.19	−0.44	−0.04	−0.03	−0.00	−0.01
	(0.14)	(0.46)	(0.21)	(0.03)	(0.02)	(0.02)
Basu, Fernald, and Kimball (2006)	0.04	0.25	0.06	0.00	0.00	0.01
	(0.07)	(0.13)	(0.08)	(0.01)	(0.01)	(0.01)
Gilchrist and Zakrajsek (2012)	−0.33	−1.22	−0.58	−0.09	−0.04	−0.06
	(0.18)	(0.84)	(0.38)	(0.06)	(0.04)	(0.04)
Observations	102	60	64	60	60	60
R^2	0.17	0.27	0.09	0.21	0.06	0.17

Note: The table shows results of regressions of alternative instruments for monetary policy shocks (in columns) on structural shocks proposed in the literature for uncertainty, oil markets, fiscal policy, productivity, and financial frictions (in rows). The frequency is monthly in panel (a) and quarterly in panel (b). HAC standard errors are in parenthesis. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% level, respectively. All regressions include a constant and month/quarter dummies. The monetary instruments are RR04—Romer and Romer (2004), BBE05—Bernanke, Boivin, and Eliasz (2005), BC13—Barakchian and Crowe (2013), GK15—Gertler and Karadi (2015), MR18—Miranda-Agrippino and Ricco (2018), and the first factor of these.

the federal funds rate and consumption expenditure prices. However, the two uncertainty measures have puzzling opposite signs, potentially reflecting multicollinearity.

Focusing on the instruments that have been identified as valid instruments by the LR-tests, Figure 4 compares the implied effects of a monetary policy shock on the endogenous variables. The solid lines show the estimates using the narrative-based measure together with 95% confidence bands for the comparison. The other proxies imply similar effects. The responses are not distinguishable from those implied by the narrative-based measure according to the confidence bands of the latter. This suggests that the estimates of the previous section provide a reasonable description of the effects of monetary policy. Moreover, it reflects the finding that a standard heteroskedas-



Notes: The figure shows the impulse responses to a 100 basis points monetary policy shock. The sample is 1980M1–2007M6 and the different instruments, using one at a time, are the narrative measure of Romer and Romer (2004) (solid line with shaded 95% pointwise confidence intervals based on 5,000 bootstrap replications), the high-frequency proxy of Gertler and Karadi (2015) (dotted lines), the model-based measure of Bernanke et al. (2005) (long dashed lines), the high-frequency cleaned instrument of Miranda-Agrippino and Ricco (2018), and the first factor (short dashed lines).

FIGURE 4. Responses for heteroskedastic proxy-SVAR using different instruments.

tic SVAR and the heteroskedastic proxy-SVAR estimate similar effects. The identification through time-varying volatility seems to dominate the information in the instruments considered here.

We conclude that most of the instruments for monetary policy shocks proposed in the literature are valid. In particular, our findings support the estimates of Romer and Romer (2004), Bernanke, Boivin, and Eliasz (2005), and Miranda-Agrippino and Ricco (2018). In contrast, they indicate some problems of the higher frequency instruments, in line with Ramey (2016) who documents partially puzzling effects of monetary policy shocks identified using these instruments.

5. CONCLUSIONS

We propose a structural vector autoregressive framework that combines the information contained in external instruments and in time-varying second moments of the data for identification of latent monetary policy shocks in the U.S. We show that the framework sharpens structural inference. Moreover, it allows testing the validity, that is, both the exogeneity and relevance, of instruments using likelihood ratio tests. Finally, it facilitates

an economic interpretation of the structural shock of interest, which is not only identified statistically through heteroskedasticity, but also through prior economic reasoning contained in the instrument. These three features of the encompassing model increase the reliability of the estimation results.

We apply the framework to test the narrative measure of monetary surprises of Romer and Romer (2004). We find that it is a valid instrument for monetary policy shocks. Using it in combination with the heteroskedasticity in the data, we provide new and potentially sharper estimates of the dynamic effects of monetary policy on the macroeconomy. We find that a surprise monetary contraction of 100 basis points in the federal funds rate leads to a significant decline in economic activity and prices of 1.6% and 0.8%, respectively. In contrast, a standard proxy-SVAR that does not exploit time-varying volatility implies substantially smaller effects.

Finally, we evaluate different proxies for monetary policy proposed in the literature. We find that instruments based on intradaily data that are further cleansed (Miranda-Agrippino and Ricco (2018)) and instruments from time-series models (Bernanke, Boivin, and Eliasz (2005)) are also valid. In our framework, they lead to qualitatively and quantitatively similar results as the narrative-based proxy.

References

Amir-Ahmadi, Pooyan, Christian Matthes, and Mu-Chun Wang (2016), "Drifts and volatilities under measurement error: Assessing monetary policy shocks over the last century." *Quantitative Economics*, 7 (2), 591–611. [163, 165, 187]

An, Sungbae and Frank Schorfheide (2007), "Bayesian analysis of DSGE models." *Econometric Reviews*, 26 (2–4), 113–172. [170]

Angelini, Giovanni and Luca Fanelli (2018), "Identification and estimation issues in structural vector autoregressions with external instruments." [168]

Antolín-Díaz, Juan and Juan F. Rubio-Ramírez (2018), "Narrative sign restrictions for svars." *American Economic Review*, 108 (10), 2802–2829. [164, 187]

Arias, Jonas E., Juan F. Rubio-Ramirez, and Daniel F. Waggoner (2021), "Inference in Bayesian proxy-svars." *Journal of Econometrics*. [164]

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016), "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131 (4), 1593–1636. [193, 194]

Barakchian, S. Mahdi and Christopher Crowe (2013), "Monetary policy matters: Evidence from new shocks data." *Journal of Monetary Economics*, 60 (8), 950–966. [163, 191, 192, 193, 194]

Basu, Susanto, John G. Fernald, and Miles S. Kimball (2006), "Are technology improvements contractionary?" *American Economic Review*, 96 (5), 1418–1448. [193, 194]

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz (2005), "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach." *The Quarterly Journal of Economics*, 120 (1), 387–422. [163, 191, 192, 193, 194, 195, 196]

Bernanke, Ben S. and Ilian Mihov (1998), "Measuring monetary policy." *The Quarterly Journal of Economics*, 113 (3), 869–902. [187]

Bertsche, Dominik and Robin Braun (2020), "Identification of structural vector autoregressions by stochastic volatility." *Journal of Business & Economic Statistics*, 0 (0), 1–14. [164]

Caldara, Dario and Edward Herbst (2019), "Monetary policy, real activity, and credit spreads: Evidence from Bayesian proxy SVARs." *American Economic Journal: Macroeconomics*, 11 (1), 157–192. [162, 165, 167, 188]

Cesa-Bianchi, Ambrogio, Gregory Thwaites, and Alejandro Vicondoa (2016), "Monetary policy transmission in an open economy: New data and evidence from the United Kingdom." Report, London, School of Economics and Political Science, LSE Library. [162, 168, 192]

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans (1999), "Monetary policy shocks: What have we learned and to what end?" *Handbook of Macroeconomics* 1, 65–148. [183, 188]

Coibion, Olivier (2012), "Are the effects of monetary policy shocks big or small?" *American Economic Journal: Macroeconomics*, 4 (2), 1–32. [188]

Fisher, Jonas D. M. and Ryan Peters (2010), "Using stock returns to identify government spending shocks." *The Economic Journal*, 120 (544), 414–436. [193, 194]

Gertler, Mark and Peter Karadi (2015), "Monetary policy surprises, credit costs, and economic activity." *American Economic Journal: Macroeconomics*, 7 (1), 44–76. [162, 163, 167, 171, 188, 191, 192, 193, 194]

Gilchrist, Simon and Egon Zakrajsek (2012), "Credit spreads and business cycle fluctuations." *American Economic Review*, 102 (4), 1692–1720. [193, 194]

Hamilton, James D. (2003), "What is an oil shock?" *Journal of Econometrics*, 113 (2), 363–398. [193, 194]

Hansen, Bruce E. (1992), "The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP." *Journal of Applied Econometrics*, 7(S1). [184]

Herwartz, Helmut and Helmut Lütkepohl (2014), "Structural vector autoregressions with Markov switching: Combining conventional with statistical identification of shocks." *Journal of Econometrics*, 183 (1), 104–116. [162, 163, 169, 184, 187, 188]

Jentsch, Carsten and Kurt G. Lunsford (2016), "Proxy svars: Asymptotic theory, bootstrap inference, and the effects of income tax changes in the United States." *American Economic Review*. [169]

Justiniano, Alejandro and Giorgio E. Primiceri (2008), "The time-varying volatility of macroeconomic fluctuations." *American Economic Review*, 98 (3), 604–641. [163, 165]

Kilian, Lutz (2008), "Exogenous oil supply shocks: How big are they and how much do they matter for the us economy?" *The Review of Economics and Statistics*, 90 (2), 216–240. [193, 194]

Lanne, Markku and Helmut Lütkepohl (2008), "Identifying monetary policy shocks via changes in volatility." *Journal of Money, Credit and Banking*, 40 (6), 1131–1149. [162, 165, 166, 167, 191]

Lanne, Markku, Helmut Lütkepohl, and Katarzyna Maciejowska (2010), "Structural vector autoregressions with Markov switching." *Journal of Economic Dynamics and Control*, 34 (2), 121–131. [166, 167, 184]

Leeper, Eric M. (1997), "Narrative and var approaches to monetary policy: Common identification problems." *Journal of Monetary Economics*, 40 (3), 641–657. [163, 188]

Leeper, Eric M. and Tao Zha (2003), "Modest policy interventions." *Journal of Monetary Economics*, 50 (8), 1673–1700. [187]

Ludvigson, Sydney C., Sai Ma, and Serena Ng (2017), "Shock restricted structural vector-autoregressions." Technical report, National Bureau of Economic Research. [164]

Lütkepohl, Helmut and Thore Schlaak (2018), "Choosing between different time-varying volatility models for structural vector autoregressive analysis." *Oxford Bulletin of Economics and Statistics*, 80 (4), 715–735. [170, 184]

Lütkepohl, Helmut and Thore Schlaak (2019), "Bootstrapping impulse responses of structural vector autoregressive models identified through GARCH." *Journal of Economic Dynamics and Control*, 101, 41–61. [169]

Lütkepohl, Helmut and Thore Schlaak (forthcoming), "Heteroskedastic proxy vector autoregressions." *Journal of Business & Economic Statistics*, 1–36. [171]

Lütkepohl, Helmut and Tomasz Woźniak (2020), "Bayesian inference for structural vector autoregressions identified by Markov-switching heteroskedasticity." *Journal of Economic Dynamics and Control*, 113, 103862. [184, 191]

Mertens, Karel and Morten O. Ravn (2013), "The dynamic effects of personal and corporate income tax changes in the United States." *American Economic Review*, 103 (4), 1212–1247. [162, 163, 165, 167, 171, 174]

Miranda-Agrippino, Silvia and Giovanni Ricco (2018), *The Transmission of Monetary Policy Shocks.* Technical report, CEPR Discussion Papers. [162, 163, 167, 191, 192, 193, 194, 195, 196]

Nakamura, Emi and Jón Steinsson (2018), "High-frequency identification of monetary non-neutrality: The information effect." *The Quarterly Journal of Economics*, 133 (3), 1283–1330. [162, 187]

Normandin, Michel and Louis Phaneuf (2004), "Monetary policy shocks: Testing identification conditions under time-varying conditional volatility." *Journal of Monetary Economics*, 51 (6), 1217–1243. [162, 165, 166, 187, 191]

Owyang, Michael T. and Garey Ramey (2004), "Regime switching and monetary policy measurement." *Journal of Monetary Economics*, 51 (8), 1577–1597. [187]

Plagborg-Møller, Mikkel and Christian K. Wolf (2021), "Local projections and vars estimate the same impulse responses." *Econometrica*, 89 (2), 955–980. [172]

Podstawski, Maximilian and Anton Velinov (2018), "The state dependent impact of bank exposure on sovereign risk." *Journal of Banking & Finance*, 88, 63–75. [169, 184]

Primiceri, Giorgio E. (2005), "Time varying structural vector autoregressions and monetary policy." *The Review of Economic Studies*, 72 (3), 821–852. [187]

Psaradakis, Zacharias and Nicola Spagnolo (2006), "Joint determination of the state dimension and autoregressive order for models with Markov regime switching." *Journal of Time Series Analysis*, 27 (5), 753–766. [183]

Ramey, Valerie A. (2011), "Identifying government spending shocks: It's all in the timing." *The Quarterly Journal of Economics*, 126 (1), 1–50. [193, 194]

Ramey, Valerie A. (2016), "Macroeconomic shocks and their propagation." In *Handbook of Macroeconomics*, Vol. 2, 71–162, Elsevier. [188, 189, 195]

Ramey, Valerie A. and Daniel J. Vine (2011), "Oil, automobiles, and the us economy: How much have things really changed?" *NBER Macroeconomics Annual*, 25 (1), 333–368. [193, 194]

Rey, Hélène (2016), "International channels of transmission of monetary policy and the mundellian trilemma." *IMF Economic Review*, 64 (1), 6–35. [163, 188]

Rigobon, Roberto (2003), "Identification through heteroskedasticity." *Review of Economics and Statistics*, 85 (4), 777–792. [162, 185, 187]

Rigobon, Roberto and Brian Sack (2003), "Measuring the reaction of monetary policy to the stock market." *The Quarterly Journal of Economics*, 118 (2), 639–669. [163, 187]

Rigobon, Roberto and Brian Sack (2004), "The impact of monetary policy on asset prices." *Journal of Monetary Economics*, 51 (8), 1553–1575. [162, 165, 166]

Rogers, John H., Chiara Scotti, and Jonathan H. Wright (2018), "Unconventional monetary policy and international risk premia." *Journal of Money, Credit and Banking*, 50 (8), 1827–1850. [162]

Romer, Christina D. and David H. Romer (1989), "Does monetary policy matter? A new test in the spirit of Friedman and Schwartz." *NBER Macroeconomics Annual*, 4, 121–170. [188]

Romer, Christina D. and David H. Romer (2004), "A new measure of monetary shocks: Derivation and implications." *American Economic Review*, 94 (4), 1055–1084. [163, 171, 183, 187, 188, 192, 193, 194, 195, 196]

Romer, Christina D. and David H. Romer (2010), "The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks." *American Economic Review*, 100 (3), 763–801. [193, 194]

Schlaak, Thore, Malte Rieth, Maximilian Podstawski (2023), "Supplement to 'Monetary policy, external instruments, and heteroskedasticity'." *Quantitative Economics Supplemental Material*, 14, https://doi.org/10.3982/QE1511. [164]

Sentana, Enrique and Gabriele Fiorentini (2001), "Identification, estimation and testing of conditionally heteroskedastic factor models." *Journal of Econometrics*, 102 (2), 143–164. [187]

Sims, Christopher A. and Tao Zha (2006), "Were there regime switches in us monetary policy?" *American Economic Review*, 96 (1), 54–81. [187]

Smets, Frank and Rafael Wouters (2007), "Shocks and frictions in us business cycles: A Bayesian dsge approach." *American Economic Review*, 97 (3), 586–606. [193, 194]

Stock, James H. and Mark W. Watson (2002), "Has the business cycle changed and why?" *NBER Macroeconomics Annual*, 17, 159–218. [163, 165]

Stock, James H. and Mark W. Watson (2012), "Disentangling the channels of the 2007-09 recession." *Brookings Papers on Economic Activity*, 120–157. [163, 165, 167, 174, 188, 193]

Stock, James H. and Mark W. Watson (2018), "Identification and estimation of dynamic causal effects in macroeconomics using external instruments." *The Economic Journal*, 128 (610), 917–948. [162]

Stock, James H., Jonathan H. Wright, and Motohiro Yogo (2002), "A survey of weak instruments and weak identification in generalized method of moments." *Journal of Business & Economic Statistics*, 20 (4), 518–529. [163, 174]

Tenreyro, Silvana and Gregory Thwaites (2016), "Pushing on a string: Us monetary policy is less powerful in recessions." *American Economic Journal: Macroeconomics*, 8 (4), 43–74. [163, 188]

Wieland, Johannes F. and Mu-Jeung Yang (2016), "Financial dampening." NBER Working Papers 22141, National Bureau of Economic Research, Inc. [183]

Wright, Jonathan H. (2012), "What does monetary policy do to long-term interest rates at the zero lower bound?" *The Economic Journal*, 122 (564), 447–466. [162]

Co-editor Tao Zha handled this manuscript.

Manuscript received 13 December, 2019; final version accepted 7 January, 2022; available online 16 February, 2022.