Ehrenbergerova, Dominika; Bajzik, Josef; Havranek, Tomas

Working Paper

When Does Monetary Policy Sway House Prices? A Meta-Analysis

When Does Monetary Policy Sway House Prices? A Meta-Analysis*

Dominika Ehrenbergerova[a,b], Josef Bajzik[a,b], Tomas Havranek[b,c]

[a]Czech National Bank
[b]Charles University, Prague
[c]CEPR

May 26, 2021

Abstract

Several central banks have leaned against the wind in the housing market by increasing the policy rate preemptively to prevent a bubble. Yet the empirical literature provides mixed results on the impact of short-term interest rates on house prices: the estimated semi-elasticities range from −12 to positive values. To assign a pattern to these differences, we collect 1,447 estimates from 31 individual studies that cover 45 countries and 69 years. We then relate the estimates to 39 characteristics of the financial system, business cycle, and estimation approach. Our main results are threefold. First, the mean reported estimate is exaggerated by publication bias, because insignificant results are underreported. Second, omission of important variables (liquidity and long-term rates) likewise exaggerates the effects of short-term rates on house prices. Third, the effects are stronger in countries with more developed mortgage markets and generally later in the cycle when the yield curve is flat and house prices enter an upward spiral.

Keywords: Interest rates, house prices, monetary policy transmission, meta-analysis, publication bias, Bayesian model averaging

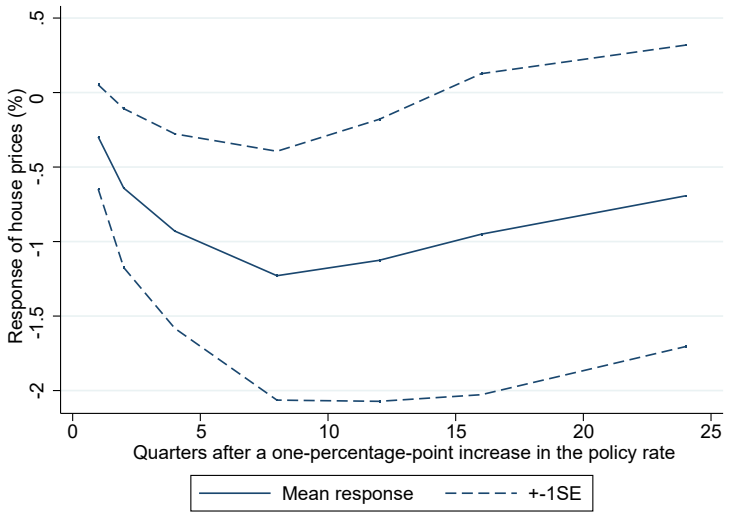
JEL Codes: C83, E52, R21

1 Introduction

Common wisdom has it that monetary policy is largely responsible for asset bubbles, including the rising house prices. That view sometimes translates into policy, such as in the case of the Swedish Riksbank between 2010 and 2014 or the government of New Zealand in 2021. In the most famous example of leaning against the wind, the Riksbank increased its policy rate from near zero to 2% in order to tame household indebtedness and house prices, even at substantial

*An online appendix with data and code is available at `meta-analysis.cz/house_prices`. Corresponding author: Dominika Ehrenbergerova, `dominika.ehrenbergerova@cnb.cz`.

Figure 1: Mean reported response of house prices to a monetary tightening



Notes: Computed based on 1,447 estimates from 221 impulse responses reported in 31 papers.

costs in terms of inflation and unemployment (Svensson, 2014, 2017). The government of New Zealand, in turn, recently amended the mandate of the Reserve Bank of New Zealand and instructed it to consider house prices when making monetary policy decisions (Powell & Wessel, 2021). The policy change in New Zealand is interesting both because its Reserve Bank has been an influential pioneer of innovations in central banking (introducing inflation targeting in 1990) and because by 2021 a large amount of research has amassed on the effects of monetary policy on house prices. This recent research, however, is rarely cited in the policy debate, which remains influenced by the arguments of Taylor (2007) in favor of the effectiveness of short-term rates in taming bubbles. Perhaps one of the reasons for the relatively limited impact of the recent research is the variance in results. The literature lacks a synthesis that would assign a pattern to the different conclusions. That is what we attempt to provide in this paper.

Figure 1 shows the mean response of house prices to a one-percentage-point increase in the short-term monetary policy rate. The mean is extracted from 221 impulse responses reported in 31 studies. The impulse responses, computed from vector autoregressions (VARs, Sims, 1980), are the main output of these studies. Hence our meta-analysis is unusual in that we collect and examine graphical results: the exact numerical results are rarely reported. For selected time horizons after the monetary policy shock we carefully measure pixel coordinates and collect the estimated response of house prices. Meta-analyses of graphical results are rare, and a prominent recent example is the meticulous survey by Fabo *et al.* (2021) on the effects of quantitative easing. Note that Figure 1 shows the corresponding 68% confidence interval (one standard error on both sides of the mean), which is the norm in the VAR literature. (Few impulse responses would be statistically significant at the 5% level common in most other fields of economics.) The impulse response bottoms out after two years at a 1.2% decrease in house prices following a one-percentage-point increase in the policy rate. We will call this effect, here 1.2, a semi-elasticity. It is clear that, on average, with such a small semi-elasticity it is

implausible for central banks to combat double-digit inflation in house prices. The implication echoes Williams (2016), a concise narrative survey of 10 earlier papers on the topic.

But a mean figure conceals important differences in the context in which the impulse response is estimated. Perhaps in some countries and certain phases of the business cycle, leaning against the wind can help moderate the increase in house prices (and, vice versa, a loose policy may help reflate depressed housing markets). Calza *et al.* (2013) suggest that the transmission of monetary policy to house prices is stronger in countries with larger flexibility and development of mortgage markets. Similarly, Iacoviello & Minetti (2003) show that financial liberalization can be important for the strength of transmission. Assenmacher-Wesche & Gerlach (2010) examine whether transmission differs between boom and standard periods. Or perhaps the small mean response is contaminated by measurement problems, such as simple recursive identification (stressed, for example, by Bjørnland & Jacobsen, 2010) and omission of important variables, such as credit (Assenmacher-Wesche & Gerlach, 2010). Our comparative advantage to the studies mentioned above is the richness of the meta-analysis dataset. No previous study in this literature has used data for more than 19 countries, which has made it difficult to investigate cross-country differences. Few cross-country studies examine more than a couple of business cycles. Similarly, comparisons of results with different identification of VAR models within individual studies have so far lacked statistical power. The work of the researchers who have collectively produced 221 impulse responses for various contexts allows us to examine the heterogeneity in transmission systematically.

Another problem with the mean impulse response is potential publication bias (Stanley, 2001),[1] which stems from the selective reporting of results that have the intuitive sign or are statistically significant. Vector autoregressions are complex models with (at least in this literature) typically few degrees of freedom. It follows that the resulting impulse response is sometimes counterintuitive: for example, it can show that house prices do not react to policy rates, or even more puzzlingly that house prices rise following a monetary tightening. If researchers take such results as evidence that their model is misspecified, they can try to run different specifications until they obtain the desired outcome. The problem is that while the puzzling impulse responses can indeed arise because of misspecifications, they can also appear simply by chance, especially given the small datasets in the literature. Seemingly large estimated effects of monetary policy in the right direction can also be due to misspecifications or chance, but it is difficult to identify them. Zero is a clear psychological cutoff that is not mirrored by a corresponding upper threshold and thereby causes a bias towards larger effects. Correction for such publication bias is thus another contribution that a meta-analysis brings on top of the results of primary studies.

Publication bias does not imply cheating and is inevitable in observational empirical research even if all researchers are honest. (In experimental research the bias can potentially be tackled by

[1]For recent papers on publication bias in economics, including positive and negative evidence, see Havranek (2015), Brodeur *et al.* (2016), Bruns & Ioannidis (2016), Ioannidis *et al.* (2017), Card *et al.* (2018), Christensen & Miguel (2018), DellaVigna *et al.* (2019), Blanco-Perez & Brodeur (2020), Brodeur *et al.* (2020), and Imai *et al.* (2021). Earlier influential papers on publication bias include Card & Krueger (1995), Ashenfelter *et al.* (1999), Ashenfelter & Greenstone (2004), Stanley (2005), and Stanley (2008).

the preregistration of experiments, see, for example, Olken, 2015, but preregistration is difficult when data are publicly available, so that the researcher can inspect them before preregistration). Publication selection can even improve the results of individual studies. The underlying effect of policy rate hikes on house prices will most likely be negative in most if not all contexts, so it is likely that the "wrong" sign indeed suggests to a researcher a problem with specification, sample size, or both. Thus it will improve the conclusions of an individual study when it does not focus on positive or zero responses of house prices. The idea of sign restrictions in vector autoregressions, eloquently advocated by Uhlig (2005), builds on a related principle. The problem is that under selective reporting the literature becomes biased as a whole since large estimates, also given by chance or misspecifications, are rarely omitted. So with individual studies we never know how much they suffer from publication bias.

For the basic identification of publication bias correction techniques we use the analogy suggested by McCloskey & Ziliak (2019), who compare publication selection to the Lombard effect in psychoacoustics: speakers involuntarily increase their vocal effort with increasing background noise. Similarly, given the example in the previous paragraph, many researchers will try harder to change the specification of their vector autoregression model if they have small samples and thus a lot of noise in estimation, a noise that often leads to insignificant initial estimates. With sufficient effort, the VAR model can be adjusted in a way that it produces point estimates large enough to outweigh the large standard errors and thus delivers statistical significance. Therefore, selective reporting creates a correlation between estimates and standard errors, a correlation that otherwise should not appear in the literature. Aside from linear tests based on the Lombard effect (regressions of estimates on standard errors) we also employ recently developed nonlinear techniques by Andrews & Kasy (2019), Furukawa (2021), and van Aert & van Assen (2021). The latter technique, p-uniform*, relaxes the assumption of no correlation between estimates and standard errors in the absence of publication bias; the assumption is perhaps too strong for the VAR literature where the impulse responses are nonlinear combinations of underlying (unreported) regression coefficients. All techniques agree that the exaggeration due to publication bias is at least twofold.

In the second part of the analysis we relate the estimated impulse responses to the context in which they were estimated. To this end we collect 39 variables that reflect the characteristics of data (e.g. time coverage), specification (e.g. inclusion of long-term interest rates), estimation (e.g. nonrecursive identification), publication (e.g. the number of citations per year), and countries (e.g. the mean share of mortgages with a floating rate in the period for which the impulse response was estimated). To tackle model uncertainty in relating the estimated semi-elasticities to the 39 explanatory variables we employ Bayesian (Raftery *et al.*, 1997; Eicher *et al.*, 2011; Steel, 2020) and frequentist (Hansen, 2007; Amini & Parmeter, 2012) model averaging. We address collinearity by using the dilution prior (George, 2010). The finding of substantial publication bias is robust to controlling for heterogeneity. Regarding data characteristics, our results suggest that studies covering shorter time series tend to produce stronger responses of house prices to monetary shocks (that is, larger semi-elasticities in the absolute

value), which is consistent with a small-sample bias. Regarding specification characteristics, we find that the omission of long-term interest rates and variables related to liquidity (credit or money supply) is associated with stronger responses of house prices to changes in the policy rate. The omitted variable bias is substantial and can strengthen the reported semi-elasticity by one percentage point. In contrast, we find little evidence that estimation and publication characteristics help explain the heterogeneity observed in the literature.

The factors most useful in explaining the differences in impulse responses are variables reflecting structural heterogeneity: the characteristics of the countries and periods for which the impulse responses were produced. Three variables are especially important. First, it is the degree of development of the mortgage market (and credit markets in general). With larger credit markets in relation to GDP, the transmission of monetary policy to house prices gets stronger. Second, it is the slope of the yield curve. With flatter yield curves, the reported semi-elasticities are larger in the absolute value. Third, it is the period of a prolonged rise in house prices: when house prices have increased for several years in a row, monetary policy becomes more potent at taming them. These country- and time-level characteristics can alter the implied impulse response by up to three percentage points. Therefore while on average house prices do not respond much to monetary policy, policy rates can help alleviate the build-up of housing bubbles in countries with developed mortgage markets during the latter part of the business cycle. Such alleviation is nevertheless costly in terms of inflation and unemployment, because even the most optimistic estimates implied by our analysis for outlying countries and time periods suggests that, after correction for publication bias, a one-percentage-point increase in the policy rate is associated with a decrease in house prices of less than 3%.

2 The Semi-Elasticity Dataset

We collect estimates of the effect of changes in the policy rate on house prices. In general, these estimates are produced in the modern literature by two types of models: dynamic stochastic general equilibrium (DSGE) models and vector autoregression (VAR) models. The results of both can be interpreted as empirical estimates, though always conditional on theoretical considerations. DSGE models need to be calibrated (or their priors set), and of course their structure is entirely based on theory. The identification of VAR models, in turn, often has theoretical foundations as well, but in some cases only as a afterthought. Compared to DSGE, VAR models are generally more data-driven, and the corresponding estimates are better suited for meta-analysis methods. Moreover, DSGE estimates of the semi-elasticity are relatively rare (a prominent example being Iacoviello & Neri, 2010). To avoid comparing apples and oranges, we focus on VAR estimates only. A general structural VAR model has the following form:

$$A_0^i Y_t^i = a^i + \gamma^i t + A^i(L) Y_{t-1}^i + B^i(L) z_t + e_t^i, \tag{1}$$

where Y_t^i is a vector of endogenous variables (including policy rates and house prices) for time t and country i, a^i is a constant, $A^i(L)$ and $B^i(L)$ are distributed lag polynomials, z_t is a

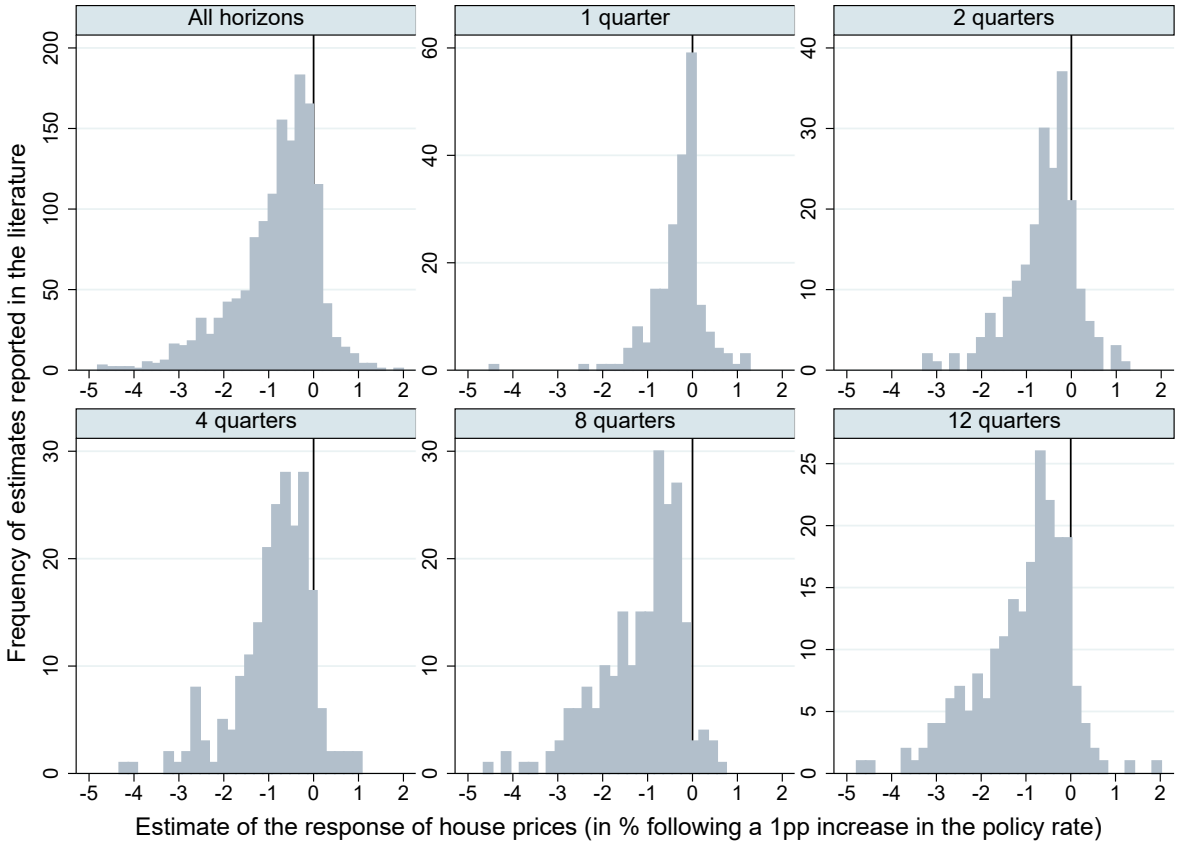
vector of exogenous variables, and e_t^i is an error term. The set of endogenous variables in a relevant VAR model usually includes output in addition to short-term rates and house prices. Depending on model specification, it may also include other variables, such the exchange rate, consumption, money supply, long-term interest rates, residential investment, and credit. In order to estimate (1), researchers rewrite it in reduced form. The principal outputs from VAR models, the reactions of the endogenous variables to structural shocks, are usually reported graphically as impulse response functions, which are easy for the reader to interpret and which cover the response over several time horizons.

To search for relevant studies we use Google Scholar because of its catholic coverage and full-text capabilities. (More details on our search strategy, including the exact query, are available in Appendix A.) We calibrate our search query in order to obtain the best known studies among the first hits. For feasibility, we only inspect the first 500 papers produced by the search. Each study needs to fulfill the following three criteria: First, for quantitative comparability the study must use a VAR model that includes house prices (not house price inflation); we thus cannot use a few influential studies such as Fratantoni & Schuh (2003) and Del Negro & Otrok (2007). Second, monetary policy must be proxied by the short-term interest rate. Third, the study must report confidence intervals around the mean impulse response function so that we can recover the precision of the estimate, which is essential for tests of publication bias. These criteria leave a total of 31 studies, which collectively use unbalanced data from 45 countries between 1947 and 2015. We add the last study in January 2021. The list of included studies, together with data and code, is available in an online appendix at `meta-analysis.cz/house_prices`.

From these 31 studies we collect the responses of house prices to a change in the policy rate after one, two, four, eight, twelve, sixteen, and twenty-four quarters. In each case we carefully measure pixel coordinates to recover the numerical estimate as precisely as possible. Specifically, we gather 208 and 211 responses after one and two quarters, respectively, and label these as short-term effects. To capture mid-term effects we gather 221 estimates for both the four- and eight-quarter horizons. To capture long-term effects we collect 216 estimates for the twelve-quarter horizon, 211 estimates for the sixteen-quarter horizon, and 159 estimates for the twenty-four-quarter horizon. Because many studies do not report responses at the latter horizon, in the analysis we focus on horizons up to sixteen quarters, and in particular the mid-term effects (four and eight quarters) most relevant to monetary policy. In a few cases, the responses for the short-term effects (one and two quarters) are not reported as the corresponding impulse responses start at the four-quarter horizon. For each impulse response we standardize the effects so that they correspond to a percentage response of house prices to a one-percentage-point increase in the policy rate. We compute the standard error from the reported confidence intervals; in the few cases when the confidence intervals are asymmetrical, we approximate the standard errors by taking the average of both bounds.

We have already commented on the mean impulse response function, Figure 1, in the Introduction. A closer view of the distribution of semi-elasticities at different horizons is provided in Figure 2. At the one-quarter horizon, most of the estimates are close to zero, and the dis-

Figure 2: Reported effects of monetary policy on house prices at different horizons



Notes: Outliers are omitted from the graphs for ease of exposition but included in all statistical tests.

tribution is almost symmetrical. With an increasing horizon, the mass of the estimates moves to the left, and the distribution becomes asymmetrical. Note that very few estimates suggest a large response of house prices to changes in the policy rate. A couple of outliers are cut from the figure for ease of exposition (the largest one being -12%), but these are isolated cases. In total for all the horizons, 87% of all the semi-elasticities lie between -2% and 1%. Moreover, more than 50% of all the semi-elasticities lie between -1% and 0%.

In addition to the impulse response functions, we also collect 39 control variables that capture the specifics of each study in order to examine the heterogeneity in the estimates. Slightly fewer than two thirds of the variables included are collected from primary studies themselves, while the remaining third consist of external country-level variables included to examine structural heterogeneity and collected from the World Bank, OECD, and Eurostat. In accordance with the latest meta-analysis reporting guidelines (Havranek *et al.*, 2020), the data taken from individual studies (estimates, confidence intervals, and variables reflecting estimation context) were collected by two co-authors of this paper and cross-checked to eliminate potential mistakes arising from manual collection. These variables are discussed in more detail in Section 4, which focuses on the heterogeneity in the literature. In the next section we focus on publication bias, which can distort the reported semi-elasticities shown in Figure 1 and Figure 2.

3 Publication Bias

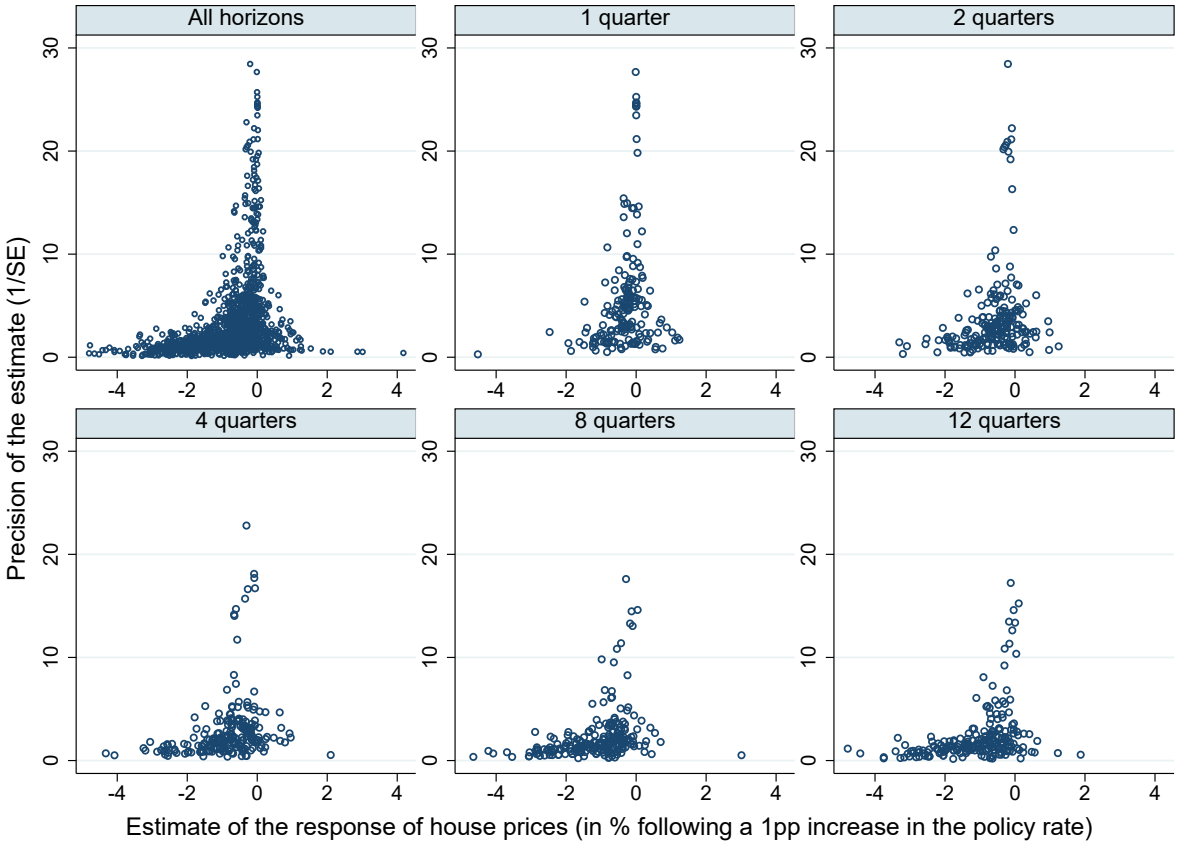
Publication bias is the systematic difference between the distribution of results produced by researchers and the distribution of results reported by researchers (both in working papers and journal articles). Sometimes the bias or its specific forms are also called selective reporting or p-hacking, though we prefer to work with the more general term. Whether publication bias is sinister or benign is still a controversial question. On the one hand, it makes little sense to build a paper on nonsensical results, such as those that suggest a rise in house prices following a monetary policy tightening. On the other hand, if such nonsensical results are ignored, the literature as a whole gets biased upwards because it is hard to spot and ignore large estimates with the right sign and significance that are also due to chance or misspecifications. The resulting tension between the effects of publication selection at the micro and macro level is in the context of vector autoregressions nicely illustrated by the following quote due to Uhlig (2012, p. 38, emphasis added):

> At a Carnegie-Rochester conference a few years back, Ben Bernanke presented an empirical paper, in which the conclusions nicely lined up with a priori reasoning about monetary policy. Christopher Sims then asked him, whether he would have presented the results, had they turned out to be at odds instead. His half-joking reply was, that he presumably would not have been invited if that had been so. There indeed is the *danger (or is it a valuable principle?)* that a priori economic theoretical biases filter the empirical evidence that can be brought to the table in the first place.

In experimental research, publication bias can in principle be eradicated by preregistration (Olken, 2015; Strømland, 2019), and the American Economic Association has established a registry for experimental papers explicitly to "counter publication bias" (Siegfried, 2012, p. 648). Such registries are also common in medical research, where publication bias has long been recognized as a grave problem (Nosek *et al.*, 2018), but we are not aware of a field in which publication bias would be extirpated by preregistration. Perhaps publication bias is allowed to survive in many fields because at the micro level of individual studies it can really represent a valuable principle, a specification check that clearly tells the researcher that something is wrong with the model or the data. It is then the task for those who evaluate the literature as a whole to correct for the macro publication bias. As we have noted in the Introduction, our basic identification procedure is based on the Lombard effect. If estimation is imprecise and data are noisy, the researcher will need to try harder to produce estimates that are fully consistent with the intuition and theory—that is, statistically significant negative estimates of house prices to a monetary tightening. So we expect more precise estimates to be less biased.

The logic of the identification assumption can be described in a so-called funnel plot often used in medical research. The funnel plot is a scatter plot of estimate size (on the horizontal axis) and estimate precision (on the vertical axis). The most precise estimates will be close to the underlying mean effect, while less precise estimates will be more dispersed, together forming the shape of an inverted funnel. If the mean underlying effect is not zero, the most precise estimates will always be statistically significant and therefore reported. In the absence of

Figure 3: Funnel plots suggest publication bias



Notes: In the absence of publication bias the scatter plots should resemble inverted funnels symmetric around the most precise estimates. Outliers are omitted from the graphs for ease of exposition but included in all statistical tests.

publication bias all imprecise estimates will be reported with the same probability. If publication bias is present and the literature as a whole prefers significant negative estimates of house prices to a monetary tightening, then given the same precision positive (and small negative) estimates will be reported with a lower probability than large negative estimates, because the latter are more likely to be statistically significant. The funnel plots reported in Figure 3 show signs of asymmetry consistent with publication bias. It is interesting to observe that the degree of asymmetry increases as the horizon of the impulse response increases, perhaps reflecting the fact that insignificant estimates are less acceptable at longer horizons.

The asymmetry of the funnel plot can be tested explicitly (Card & Krueger, 1995; Egger *et al.*, 1997):

$$\hat{x}_{i,j} = \alpha_0 + \beta SE_{i,j} + \epsilon_{i,j}, \tag{2}$$

where $\hat{x}_{i,j}$ denotes the i-th estimated effect of interest rates on house prices in the j-th study, and SE is the corresponding standard error. Parameter α_0 denotes the mean effect beyond bias (that is, conditional on infinite precision and thus no publication selection), while β represents the intensity of publication bias. The simple regression has at least two problems (aside from

ignoring heterogeneity, which we will introduce in the next section). First, it assumes a linear relationship between the standard error and the extent of publication bias. But the correlation between bias and precision can vary for different values of precision. When the estimate is very precise, small changes in precision do not alter the intensity of publication selection because they do not alter the designation of statistical significance at standard levels. When the estimate is very imprecise, small increases in precision do not achieve statistical significance and thus do not influence publication probability and selective reporting. It is for intermediate values of precision, and especially around the main threshold for statistical significance, that a relation between estimates and standard errors is more likely.

Second, (2) assumes that the standard error is exogenous. The assumption can be realistic in medical research where the standard error is basically given to the researcher (it is computed based on a straightforward formula of the number of observations), but in economics the computation of the standard error is a complex exercise. In any case the standard error is not given but can be influenced by the estimation approach; therefore publication bias can work through both point estimates and standard errors. A related problem is that the standard error itself is estimated, and thus (2) suffers from attenuation bias (Stanley, 2005).

We relax the linearity assumption by employing the stem-based method by Furukawa (2021) and the selection model by Andrews & Kasy (2019). The stem-based method (alluding to the stem of the funnel plot) is a nonparametric approach that optimizes the trade-off between bias and variance. When only the most precise studies are used to compute the mean effect, little publication bias remains, but the variance of the mean estimate increases because it is inefficient to discard information. When less precise studies are included, the variance of the mean estimate decreases, but the mean is more contaminated by bias. Furukawa (2021) presents a clever way how to weigh these two problems and select the optimal number of most precise studies for the computation of the mean effect. The selection model by Andrews & Kasy (2019) assumes that the probability of reporting for each estimate depends on its sign and statistical significance, with changes in probability at 0 and the main thresholds for statistical significance. The model then re-weights the estimates based on the computed reporting probabilities.

We relax the exogeneity assumption by employing the p-uniform* method by van Aert & van Assen (2021). The method does not assume anything about the relationship between estimates and standard errors but uses the statistical principle that the distribution of p-values should be uniform at the underlying mean effect size. Consequently, it recomputes p-values and searches for the mean value of the semi-elasticity that would be consistent with a uniform distribution of p-values. In addition, in Appendix B we use several techniques that are robust to the exogeneity assumption but do not provide estimates of the mean semi-elasticity corrected for publication bias; instead they test for the presence of publication bias (Gerber & Malhotra, 2008a; Elliott *et al.*, 2021) or test the null hypothesis that the corrected effect is zero (Simonsohn *et al.*, 2014a).

The main results are shown in Table 1. Panel A reports the findings of linear models (the regression of estimates on standard errors), while Panel B focuses on nonlinear models. We employ double clustering of standard errors at the level of studies and countries. Because we

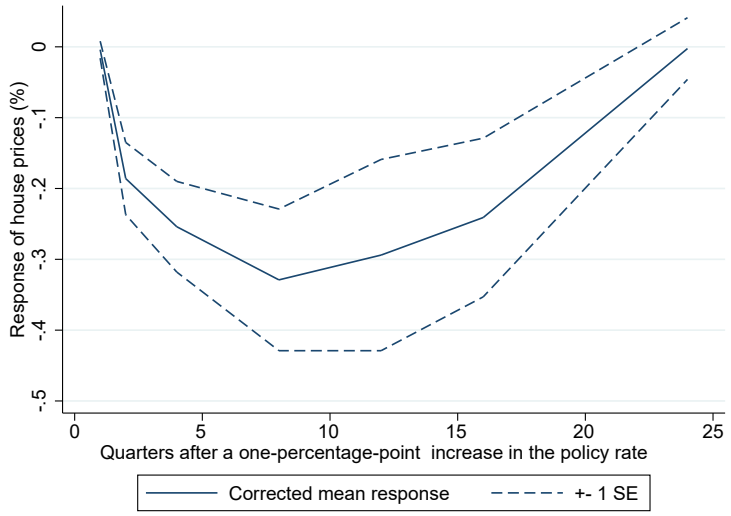
Table 1: Linear and nonlinear tests suggest publication bias

Time after a monetary policy shock:	1 quarter	2 quarters	4 quarters	8 quarters	12 quarters	16 quarters
PANEL A: Linear models						
Regression of reported estimates on their standard errors, ordinary least squares						
Standard error (publication bias)	-0.751***	-1.099***	-1.280***	-0.990***	-0.451	-0.281
	(0.238)	(0.378)	(0.456)	(0.288)	(0.281)	(0.182)
	[-1.227, -0.034]	[-1.962, -0.304]	[-2.344, -0.295]	[-2.027, -0.038]	[-1.397, 0.116]	[-0.832, 0.173]
Constant (corrected mean effect)	-0.034	-0.055	-0.094	-0.402**	-0.699***	-0.648***
	(0.074)	(0.189)	(0.256)	(0.175)	(0.202)	(0.167)
	[-0.209, 0.137]	[-0.452, 0.465]	[-0.712, 0.691]	[-0.923, 0.179]	[-1.176, -0.124]	[-1.043, -0.175]
Regression of reported estimates on their standard errors, weighted by inverse variance						
Standard error (publication bias)	-0.838***	-0.853***	-1.036***	-1.078***	-0.879***	-0.659***
	(0.165)	(0.148)	(0.204)	(0.214)	(0.250)	(0.197)
	[-1.269, -0.422]	[-1.210, -0.505]	[-1.591, -0.547]	[-1.658, -0.464]	[-1.546, -0.167]	[-1.277, -0.097]
Constant (corrected mean effect)	-0.004	-0.186***	-0.254***	-0.329***	-0.294**	-0.241**
	(0.012)	(0.051)	(0.064)	(0.100)	(0.135)	(0.112)
	[-0.0340, 0.020]	[-0.375, 0.006]	[-0.471, 0.052]	[-0.671, 0.101]	[-0.702, 0.141]	[-0.548, 0.129]
PANEL B: Nonlinear models						
Stem-based method (Furukawa, 2021)						
Corrected mean effect	-0.006	-0.208***	-0.303***	-0.324**	-0.171	-0.120
	(0.009)	(0.081)	(0.131)	(0.165)	(0.133)	(0.089)
Selection model (Andrews & Kasy, 2019)						
Corrected mean effect, break at $t = 1.645$	-0.112**	-0.190	-0.364***	-0.447***	-0.325**	-0.041
	(0.052)	(0.274)	(0.064)	(0.124)	(0.134)	(0.028)
Corrected mean effect, break at $t = 1$	0.006*	-0.121	-0.332***	-0.380***	-0.275**	-0.103
	(0.003)	(0.117)	(0.074)	(0.086)	(0.138)	(0.079)
P-uniform (van Aert & van Assen, 2021)*						
Corrected mean effect, break at $t = 1.645$	-0.181***	-0.126***	-0.144***	-0.137***	-0.122***	-0.093***
Corrected mean effect, break at $t = 1$	-0.056***	-0.091***	-0.105***	-0.107***	-0.101***	-0.087***
Observations	208	211	221	221	216	211

Notes: The mean uncorrected effect at the 8-quarter horizon was −1.2. Standard errors, clustered at the level of studies and countries, are depicted in round brackets; confidence intervals from wild bootstrap are in square brackets. The p-uniform* method reports p-values, which are all below 0.001 and thus not shown in the table. The selection model and p-uniform* require specifying the break corresponding to a publication selection rule. The wild bootstrap (Cameron *et al.*, 2008) is implemented via the *boottest* package in Stata (Roodman *et al.*, 2019). *, **, and *** denote significance at the 10%, 5%, and 1% level, respectively.

only have 31 studies in our dataset, we additionally also report confidence intervals based on wild bootstrap. In the first part of Panel A we run the regression specified in (2), while in the second part we run weighted least squares with weights proportional to the inverse variance of the reported estimates. The weighted specification corrects for the heteroskedasticity inherent in (2). In the case of the selection model and p-uniform* in Panel B we need to specify the relevant thresholds for statistical significance. As we have noted in the Introduction, it is common in the VAR literature to use the 68% confidence interval (that is, one standard error on both sides of the mean) instead of the 95% interval common elsewhere in economics. A few VAR studies use the 90% interval, so we set our thresholds for the corresponding values of the t-statistic at 1 and 1.645. Two observations emerge from the table. First, the corrected mean effect is always smaller than the simple mean. Second, the effect at the eight-quarter horizon is always statistically significant even after correction for publication bias and ranges from -0.45 (selection model) to -0.11 (p-uniform*). The presence of publication bias and significance of the mean effect corrected for publication bias is further supported by robustness checks presented in Appendix B that use the techniques of Gerber & Malhotra (2008a), Simonsohn *et al.* (2014a), and Elliott *et al.* (2021).

Figure 4: Mean impulse response after correction for publication bias



Notes: Based on the weighted least squares specification reported in Panel A of Table 1.

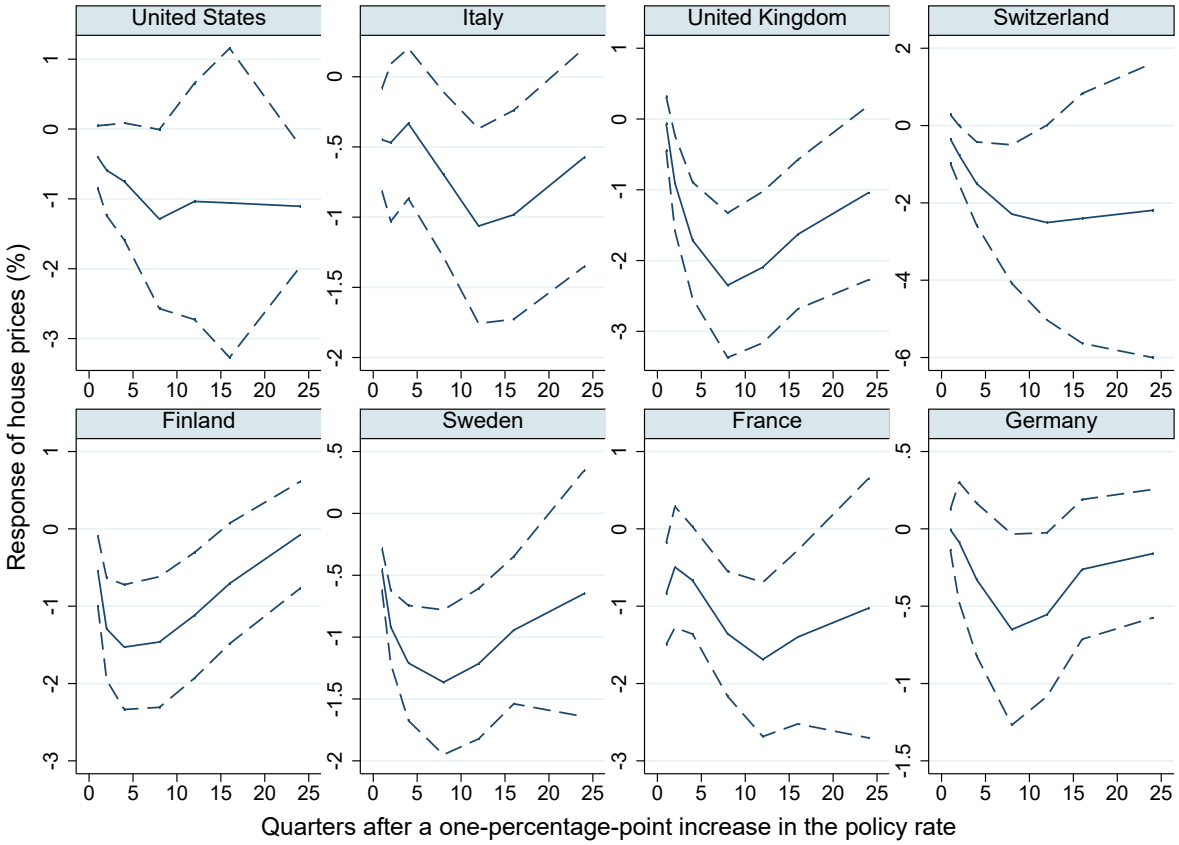
The weighted least squares specification yields estimates of the mean effect close to the median of those of all the techniques considered, and we use this specification to construct the implied impulse response corrected for publication bias. The response is shown in Figure 4 and presents a similar shape to the one discussed earlier in relation to Figure 1: house prices decrease swiftly following a monetary policy tightening, the effect peaks after two years and then dissipates. The main difference is the size of the response, which is now much smaller: -0.33% after two years compared to the simple uncorrected mean estimate of -1.2%. Publication bias thus has important quantitative implications for the estimated effectiveness of monetary policy

in taming house prices. Nevertheless, the finding of publication bias, and the mean impulse response itself, may be contaminated by the differences in the context in which the estimates are obtained. In the next section we thus turn to the heterogeneity in the estimates.

4 Heterogeneity

The previous literature has hinted on the differences in the transmission of monetary policy to house prices depending on the context of countries, time periods, and estimation techniques (among others, Iacoviello & Minetti, 2003; Assenmacher-Wesche & Gerlach, 2010; Bjørnland & Jacobsen, 2010; Calza *et al.*, 2013). But these studies could compare only a few countries, a few business cycles, and a few models computed using different specifications. Based on the efforts of these researchers, we build a large database of not only the reported results but also the factors that might have influenced those results. We are thus able to examine the heterogeneity in the response of house prices to policy rate shocks with much more power than the individual studies in the literature.

Figure 5: Cross-country heterogeneity in transmission



Notes: The graphs show mean impulse responses reported for individual countries (the solid line) and 68% confidence intervals constructed by adding one standard error to each side of the mean (the dashed lines).

Consider Figure 5, which shows mean impulse responses reported for selected countries. While all the responses are intuitive and none shows the pesky price puzzle, an increase in prices following a monetary policy tightening, the strength and speed of transmission varies greatly across countries. The maximum decrease in house prices following a one-percentage-point increase in the policy rate is -0.6% in Germany but -2.2% in the United Kingdom. In Finland house prices near their maximum response already after 2 quarters and dissipate quickly after two years, while the responses are persistent in Switzerland and the United States. The responses are quite precisely estimated for Finland, Germany, and the United Kingdom, while transmission is uncertain in France, Switzerland, and the United States. In this section we try to explain these and other differences, together with evaluating the robustness of publication bias results to controlling for heterogeneity. Aside from variables that measure the characteristics of countries and the business cycle (what we call structural heterogeneity), we also control for the characteristics of data, specification, estimation, and publication. The definitions and summary statistics for all the variables are available in Table C1 in Appendix C; the variables are also briefly summarized below. For simplicity we focus on the four-quarter horizon, which is arguably the most relevant for monetary policy if the central bank intends to defuse a housing bubble in time (the results for the eight-quarter horizon are nevertheless similar).

4.1 Variables

Data Characteristics. We control for the characteristics of the data used in the primary studies. First, regarding data frequency, only around 10% of the estimates come from studies that use monthly data; the rest are based on quarterly data. Second, we control for whether simple time series (80% of all observations) or panel data are used in vector autoregressions. We are also interested in whether the strength of transmission changes over time, and we thus include the mean year of the dataset used. By doing so, we control for the potential change in transmission not accounted for by variables capturing structural heterogeneity, which will be described below. We also test whether the length of the sample used in the primary studies systematically affect the estimates.

Specification characteristics. When assessing the effect of monetary policy on the overall price level, Rusnak *et al.* (2013) find that study design has a significant effect on the results. For instance, they find that including output gap as a measure of output or commodity prices besides overall prices systematically affects the results. In a similar way, we create dummies for additional endogenous variables included in VAR models estimating the transmission of monetary policy to house prices. We include a dummy equal to one if the GDP deflator is used instead of the usual consumer price index. Next, we include dummy variables that equal one if a measure of credit (usually real credit to the private sector or mortgage loans) is used (26% of cases), if the long-term interest rate is used (17% of cases), and if consumption, residential investment, the money supply, the exchange rate, and the foreign interest rate are included. We distinguish between nominal and real house prices, though nominal house prices are used in merely around 5% of the studies. We only include studies which use residential house prices,

not commercial house prices, land prices, or rent prices. As far as the remaining aspects of the estimation specification are concerned, we control for the number of lags included in the VAR model. The number of lags affects the persistence of the impulse responses and can thus also affect the strength of transmission.

Estimation characteristics. Another important dimension in which estimates differ is the estimation technique. The primary studies typically use a reduced-form VAR employing ordinary least squares or maximum likelihood, and they usually rely on recursive ordering as their identification scheme (77% of all estimates). We control for the use of sign restrictions. Since sign restrictions differ across papers (the restriction may not be imposed on all variables in the same direction), we distinguish between two cases that are important for the transmission to house prices. First, we include a dummy variable equal to one if sign restrictions are imposed on the house price variable, guaranteeing the expected sign. Second, we include a dummy if sign restrictions are imposed on any other variables, but not house prices. We then control for other types od nonrecursive identification (such as long-run restrictions) and, regarding the estimation procedure, we also create a dummy variable that equals one if a Bayesian VAR is estimated (around 10% of the estimates).

Publication characteristics. While the variables introduced above can help us control for some aspects of study quality, other aspects will remain difficult to code or even observe. As additional proxies for quality, we include three publication characteristics. First, we control for the number of Google Scholar citations each study has received per year on average since it appeared on Google Scholar for the first time. This way we take into account the long and variable publication lags in economics, where working papers might accumulate a significant amount of citations even prior to publication. We also include variables reflecting publication in a peer-reviewed journal and the RePEc discounted recursive impact factor of the outlet. We expect highly-cited studies published in peer-reviewed journals with a high impact factor to be of higher quality than other studies, ceteris paribus. A qualification is of course in order, because any potential correlation between the size of the estimates and the publication characteristics can be also due to publication bias and not necessarily due to genuine systematic effects of (unobserved) study quality on results. One must therefore be cautious with the interpretation of the results related to this group of variables.

Structural heterogeneity. We include a wide range of external variables (marked with the prefix "Country-level"), that is, variables obtained outside the primary studies to cover relevant macroeconomic, financial, demographic, and housing supply factors. For each impulse response, we compute these variables as mean values of the time span used to deliver the particular impulse response for a given country or a group of countries (in which case we weight the individual country-level values by country GDP). First, we include a measure of economic development—disposable income per capita. We also include separate dummy variables equal to one in boom and crisis periods. Second, we include interest rate variables, which we suspect may interact with the transmission to house prices. We control for the level of the short-term interest rate itself: transmission can be more complete at higher ("normal") monetary policy

rates, while it can change at low interest rates because of excessive risk-taking by economic agents. On the other hand, very low interest rates or prolonged periods of very low interest rates may cause asymmetries in the transmission. In consequence, a prolonged period of low interest rates fueling credit and house price booms could be mirrored by a stronger reaction of house prices to monetary policy. Long-term interest rates (10-year government bond yields) are more relevant than the short-term rates for the transmission to house prices, and they are often driven by factors independent of the policy rate, such as demographics, inequality, savings glut, the relative price of capital, and amount of public investment (Rachel & Smith, 2017). Due to collinearity concerns, we include the term premium (spread) instead the long-term rate per se. We also include the inflation rate in the country: as shown by Rusnak *et al.* (2013), periods of high inflation are often associated with a lower credibility of the central bank and thus weaker transmission.

Third, we control for the characteristics of the lending market by including the credit-to-GDP ratio in order to account for the level of indebtedness as well as for the level of financial development. The inclusion of the mortgage-to-GDP ratio yields similar results, but because the amount of mortgages is unavailable for several countries in our dataset, we use the credit-to-GDP ratio instead to increase the number of degrees of freedom available for our analysis. We also include a variable capturing the share of mortgage loans with floating interest rates: the higher the share of floating-rate mortgages, the stronger the immediate transmission to the overall mortgage interest rate, and possibly the stronger the transmission to house prices in general. For similar reasons we also control for the average maturity of mortgage loans in the country. Fourth, regarding demographic characteristics we account for population growth in the country. If population growth is high, transmission may be weaker as house prices are driven by demographics rather than being affected by monetary policy.

Fifth, we include several characteristics of the housing sector. In order to account for house supply factors, we include the number of building permits. A low number of building permits indicates restricted housing supply and potentially hampered transmission of monetary policy.[2] We also cover the home ownership structure. We include a proxy for tourism as a demand factor rather than a housing supply one. The remaining variables capturing structural heterogeneity relate to house prices themselves. In particular, we include the standardized price-to-income ratio as a proxy for overvaluation of house prices. The price-to-income ratio, available from the OECD database, is measured as the nominal house price divided by nominal disposable income per capita and can be considered a measure of affordability. As another potential proxy to capture overvalued house prices we include a variable capturing the number of periods house price growth is above its long-term average.

[2]The number of building permits acts as a proxy for housing supply. Other variables could serve this purpose: for example the number of dwellings. Nevertheless, the inclusion of this variable led to collinearity in our dataset. Another candidate variable would be an estimate of the sensitivity of house prices to housing supply. However, we are restricted by availability for our wide cross-country sample. Therefore, we stick to the number of building permits as a house supply proxy often used in the literature (e.g., Grimes & Aitken, 2010; Paciorek, 2013)

4.2 Estimation

We intend to find out whether the variables introduced above are systematically related to the reported effects of monetary policy on house prices. The easiest way would simply be to regress the estimated semi-elasticities on all the variables. But because of the large number of variables (39), such an estimation would be inefficient because many of the variables will probably not belong in the best underlying model. In other words, we face substantial model uncertainty, which is coupled by collinearity. Both can be addressed by Bayesian model averaging (BMA) with a dilution prior. BMA runs many models with different combinations of the explanatory variables and then constructs a weighted average over these models with weights being proportional to model fit and complexity. The dilution prior (George, 2010) gives each model an additional weight proportional to the determinant of the correlation matrix, so that collinearity is penalized in the final output of Bayesian model averaging. BMA was pioneered in the social sciences by Raftery (1995) and Raftery *et al.* (1997). As a robustness check, we use a frequentist alternative (frequentist model averaging, FMA), which is based on Magnus *et al.* (2010) and Amini & Parmeter (2012).

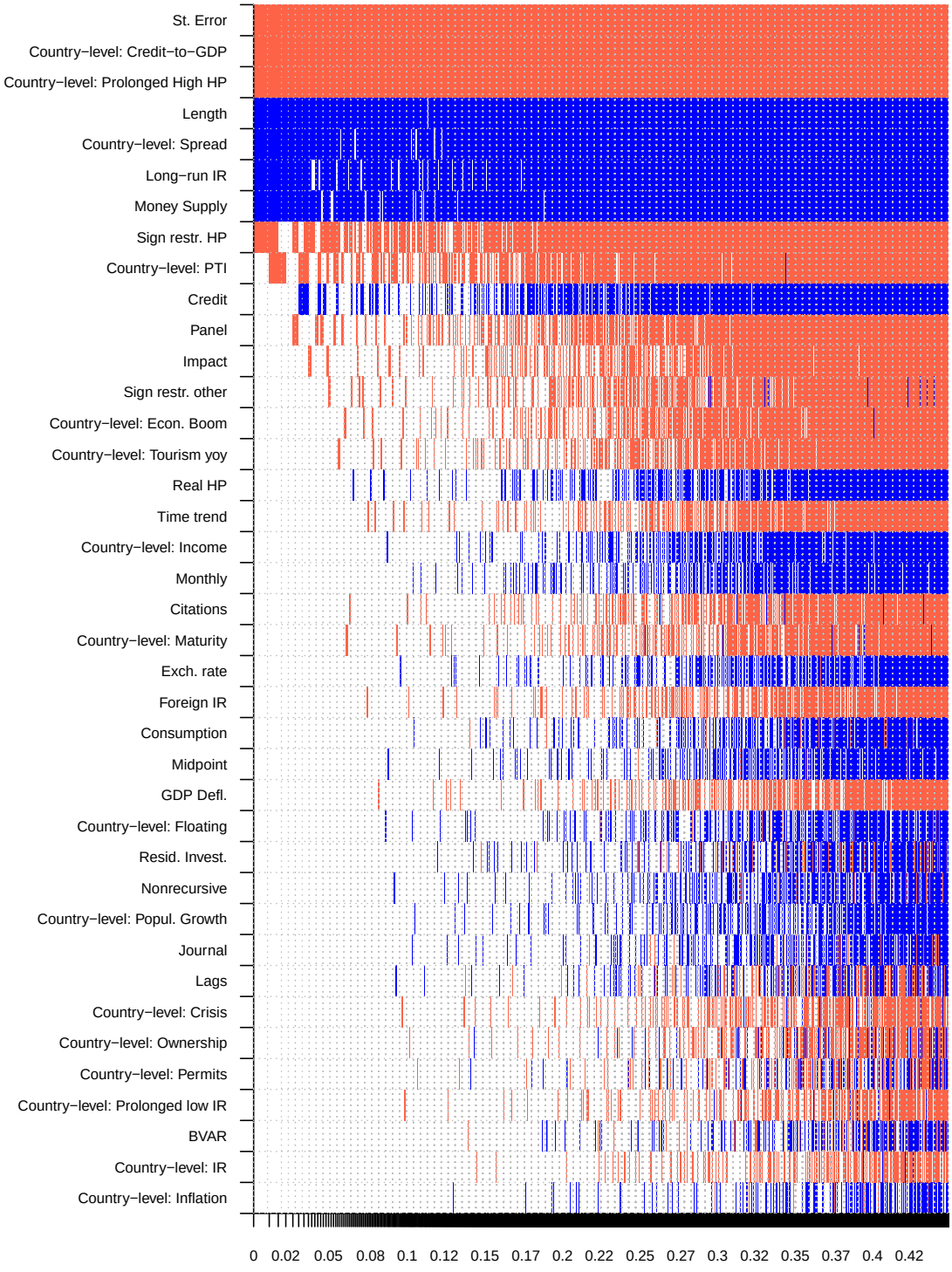
BMA can potentially run 2^{39} regressions with all the possible combinations of variables. Such a computation would take several months, and we avoid it by using the Markov Chain Monte Carlo process and its Metropolis-Hastings algorithm (Zeugner & Feldkircher, 2015), which goes through the most probable models. The posterior model probability then expresses the weight of each model. The estimated coefficients for every variable are weighted by the posterior model probability through all the models. For each variable we thus obtain a posterior inclusion probability (PIP), which denotes the sum of the posterior model probabilities of all the models in which the variable is included.

Concerning priors, in the baseline specification the unit information g-prior (UIP) recommended by Eicher *et al.* (2011) gives the prior the same weight as one observation of the data. It constitutes our benchmark setting, addressing the lack of prior knowledge regarding the parameter values. Moreover, the dilution prior addressing collinearity provides us with the benchmark model prior. Aside from the weight proportional to the determinant of the correlation matrix, all models have the same prior probability. As a robustness check of our baseline BMA results, we estimate BMA using alternative g-priors and model priors. We use a combination of the unit information g-prior and the uniform model prior and a combination of the Hannan-Quinn (HQ) g-prior and the random model prior (Fernandez *et al.*, 2001; Ley & Steel, 2009). As we have noted, we also use frequentist model averaging as an additional robustness check. In FMA we use Mallow's criterion for model averaging (Hansen, 2007), and the covariate space is orthogonalized using the approach of Amini & Parmeter (2012).

4.3 Results

Figure 6 summarizes the results of Bayesian model averaging graphically. Columns denote individual regression models from the best ones on the left, and the variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative

Figure 6: Model inclusion in Bayesian model averaging



Notes: The response variable is the estimated effect of a one-percentage-point change in the interest rate on house prices after four quarters. Columns denote individual models; the variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities; only the 10,000 best models are shown. To ensure convergence we employ 3 million iterations and 1 million burn-ins. Blue color (darker in grayscale) = the variable is included and the estimated sign is positive, i.e. transmission is weaker. Red color (lighter in grayscale) = the variable is included and the estimated sign is negative, i.e. transmission is stronger. No color = the variable is not included in the model. The numerical results of the BMA exercise are reported in Table 2. A detailed description of the variables is available in Table C1.

Table 2: Why reported impulse responses vary

Category	Variable	PIP	Post. mean	Post. SD
Publication bias	**SE**	**1.000**	**-1.510**	**0.143**
Data characteristics	Monthly	0.062	0.029	0.138
	Panel	0.212	-0.067	0.150
	Length	**0.941**	**1.511**	**0.560**
	Midpoint	0.036	0.008	0.056
Specification characteristics	GDP Defl.	0.036	-0.008	0.062
	Foreign IR	0.045	-0.014	0.104
	Credit	0.246	0.077	0.155
	Consumption	0.040	0.006	0.050
	Resid. Invest.	0.034	0.008	0.071
	Money Supply	**0.809**	**0.489**	**0.303**
	Exch. rate	0.046	0.007	0.050
	Long-run IR	**0.822**	**0.471**	**0.284**
	Real HP	0.075	0.032	0.143
	Lags	0.027	0.000	0.010
	Time trend	0.069	-0.013	0.060
Estimation characteristics	BVAR	0.016	0.003	0.056
	Sign restr. HP	**0.581**	**-0.465**	**0.460**
	Sign restr. other	0.106	-0.070	0.247
	Nonrecursive	0.031	0.004	0.052
Publication characteristics	Citations	0.057	-0.007	0.039
	Impact	0.119	-0.025	0.082
	Journal	0.028	0.002	0.029
Structural heterogeneity	Country-level: Crisis	0.026	0.000	0.004
	Country-level: IR	0.015	-0.001	0.011
	Country-level: Prolonged low IR	0.021	0.000	0.003
	Country-level: Spread	**0.872**	**0.378**	**0.200**
	Country-level: Floating	0.036	0.000	0.001
	Country-level: Tourism yoy	0.097	-0.002	0.007
	Country-level: Income	0.066	0.044	0.204
	Country-level: Inflation	0.012	0.000	0.006
	Country-level: Credit-to-GDP	**0.997**	**-0.010**	**0.003**
	Country-level: Popul. Growth	0.028	0.005	0.050
	Country-level: PTI	0.360	-0.007	0.010
	Country-level: Prolonged High HP	**0.968**	**-0.104**	**0.032**
	Country-level: Permits	0.021	0.000	0.001
	Country-level: Maturity	0.053	-0.017	0.099
	Country-level: Ownership	0.024	0.000	0.002
	Country-level: Econ. Boom	0.103	-0.005	0.016
Observations	209			

Notes: PIP = posterior inclusion probability. SD = standard deviation. Variables with a posterior inclusion probability higher than 0.5 are shown in bold. We employ the unit information g-prior as recommended by Eicher *et al.* (2011) and the dilution prior to address collinearity (George, 2010). A detailed description of the variables is available in Table C1.

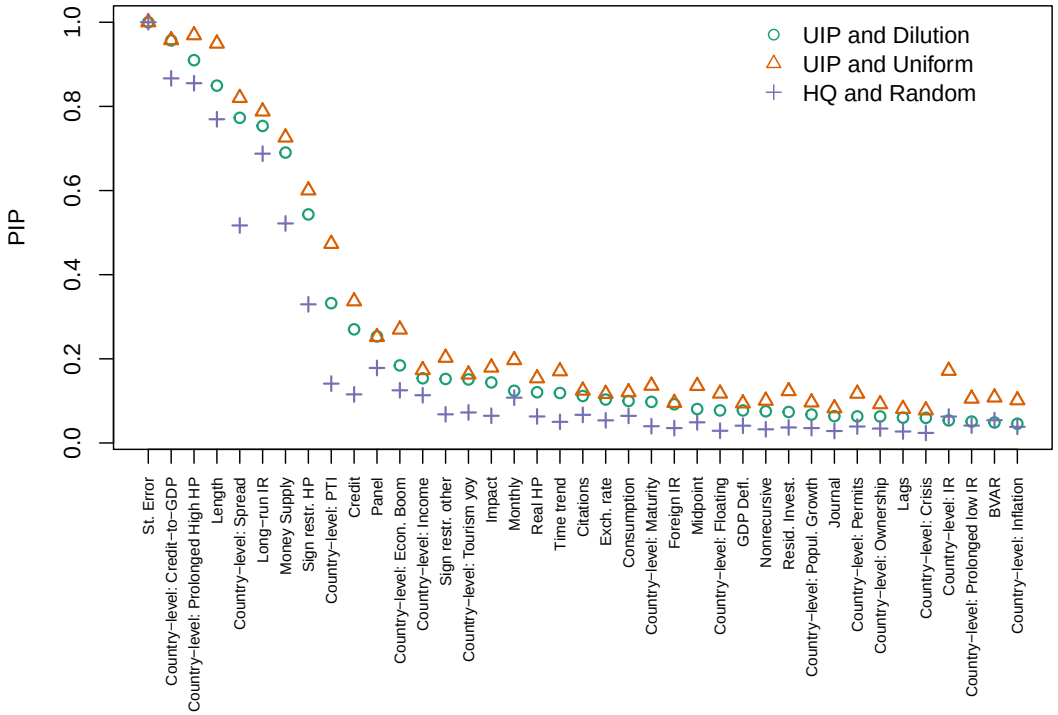
posterior model probabilities; only the 10,000 best models are shown, which is why the cumulative probability does not run to 1. To ensure convergence we employ 3 million iterations and 1 million burn-ins. Blue color (darker in grayscale) means that the variable is included and the estimated sign is positive, i.e. transmission is weaker. Red color (lighter in grayscale) means that the variable is included and the estimated sign is negative, i.e. transmission is stronger. Blank cells denote exclusion of the variable. Eight variables are included in most of the best models, which means that these variables are effective in explaining the heterogeneity in the reported semi-elasticities: the standard error (a proxy for publication bias), credit to GDP (a proxy for financial development), prolonged growth in house prices (a proxy for the build-up of a housing bubble), length of the time series (a proxy for small-sample bias), the term premium (spread between short- and long-term rates, a proxy for risk-taking and position in the business cycle), the inclusion of long-term rates and money supply (proxies for omitted variables), and sign restriction for the house prices variable (a proxy for the importance of estimation). The remaining variables have posterior inclusion probabilities below 0.5, which means they are not important in explaining the differences in reported results.

The numerical results of Bayesian model averaging are reported in Table 2. The eight variables with posterior inclusion probabilities above 0.5 are shown in bold. The posterior means presented in the table measure the partial derivatives of the reported semi-elasticities with respect to the variables in question. Our results suggest that the finding of substantial publication bias is robust to controlling for heterogeneity. Not only that the variable proves to be important in BMA, but it also has the largest posterior inclusion probability and the estimated coefficient (posterior mean) is larger than that reported in the previous section. We conclude that our previous finding of publication bias was not driven by omitting factors associated with heterogeneity. Next, we find that studies using longer time series are likely to report evidence of weaker transmission from monetary policy decisions to house prices. The result is consistent with a small-sample bias towards more negative semi-elasticities.

We find that specification characteristics are important for the reported estimates of the semi-elasticity. When long-term interest rates are omitted from the analysis, the reported response of house prices tends to be more negative. The omission of variables related to liquidity (credit or money supply) has a similar effect. Note that credit and money supply are correlated in most countries. In our baseline estimation we include both and obtain a high posterior inclusion probability for money supply and a relatively small inclusion probability for credit. If we exclude money supply from the analysis, however, the inclusion probability of credit rises above 0.5. The results highlight that house prices are affected by liquidity and long-term interest rates aside from the policy rate. We also find that studies which put a (negative) sign restriction on the response of house prices tend to find, on average, more negative effects. That finding is intuitive because with sign restrictions the price puzzle is a priori impossible. Note that sign restrictions are used only by 5% of the specifications in our sample, and the results (including those on publication bias) would not be affected by excluding these restricted estimates entirely from our analysis.

Finally, our results suggest that variables reflecting the financial system of the country and the position in the cycle are important in explaining the reported semi-elasticities. The credit-to-GDP ratio has a posterior inclusion probability of almost 1 and shows a negative correlation with the reported response of house prices. Note that the result would be similar if we used the mortgage-to-GDP ratio instead. We opt for the former because data on the amount of mortgages are not available for every country and time period of our dataset, so using mortgages would mean throwing away data. We interpret the finding, in line with Calza *et al.* (2013), as evidence for stronger transmission in countries with more developed mortgage (and, in general, credit) markets. Next, we find that a flatter yield curve and an ongoing build-up of a bubble in the housing market are both associated with stronger transmission. The result is consistent with monetary policy being more effective at influencing house prices at the latter part of the business cycle, when banks and households are more prone to excess optimism and risk-taking, and adds some credence to the policy of leaning against the wind. In the next subsection, however, we show that even under the best of circumstances the strength of transmission is insufficient to substantially mitigate housing bubbles.

Figure 7: Sensitivity to alternative priors



Notes: UIP = unit information prior; the prior has the same weight as one observation of data. Uniform model prior = each model has the same prior weight. Dilution model prior = the prior weight of each model is proportional to the determinant of the correlation matrix. The HQ prior asymptotically mimics the Hannan-Quinn criterion. The random model prior assign the same prior weight to each model size (e.g., models with 10 variables have the same prior probability as models with 11 variables). PIP = posterior inclusion probability.

As we have noted, we run several robustness checks to test the robustness of our results. Figure 7 shows the posterior inclusion probabilities for individual variables using different sets of priors in Bayesian model averaging. The changes are small and would not change our conclusions. In Appendix C we show the results of frequentist model averaging, Bayesian model averaging for all semi-elasticities (not just those at the four-quarter horizon), and ordinary least squares regressions for all horizons separately. The results of FMA are broadly consistent with those of BMA, though generally yield less significance (for example, the p-values associated with the variables reflecting the inclusion of long-term interest rates and money supply are around 0.15). On the other hand, BMA and OLS results for all semi-elasticities imply more significance for most variables compared to our baseline BMA for semi-elasticities at the four-quarter horizon. In all cases, the finding of publication bias is statistically significant at the 1% level (in frequentist techniques) or has a posterior inclusion probability of 1 (in Bayesian techniques).

4.4 Implied Response

As the bottom line of our analysis we compute the impulse response implied by the entire literature but conditional on the absence of publication bias and misspecifications. We construct both the mean impulse response for the typical country and also responses for individual countries. In general, our results can be easily used to derive an implied impulse response conditional on any selected aspect of the financial system, business cycle, and estimation techniques. Technically the implied responses are computed as fitted values using the results of Bayesian model averaging and a definition of the preferred values for each variable included in BMA (or the sample mean if no preference can be made). So we plug in zero for the standard error in order to condition the implied response on the correction for publication bias. While we have noted that the linear correction for publication bias using the exogeneity assumption for the standard error is problematic in theory, we have also shown in the previous section that in the literature on monetary transmission to house prices the linear correction gives results similar to more complex methods (if anything, it is more conservative in the correction for publication bias, perhaps due to attenuation). Since it is implausible to use the more complex methods of publication bias correction in BMA, we rely on the linear regression.

We prefer studies that cover as many years as possible, which is to say we plug in sample maximum for the variable reflecting sample length. Next, in order to put more weight on studies that use recent data we employ sample maximum for the variable capturing the mean year of data. Regarding specification characteristics, we prefer if the study uses real house prices (instead of nominal), controls for the long-term interest rate, and uses credit or money supply to control for liquidity. Regarding estimation characteristics, we prefer Bayesian techniques and nonrecursive identification (structural VAR or sign restrictions). Regarding publication characteristics, we prefer highly cited studies published in peer-reviewed journals with a high impact factor: so we plug in 1 for the dummy variable reflecting journal publication and sample maxima for the number of per-year citations and the RePEc discounted recursive impact factor of the journal. We leave all other variables, including variables capturing structural heterogeneity, at

Table 3: Best practice estimates

Horizon:	1Q	2Q	4Q	8Q	12Q	16Q
Baseline	-0.001	-0.233	-0.448	-0.678	-0.544	-0.299
Agnostic on specification	-0.737	-0.969*	-1.183**	-1.414**	-1.279**	-1.035*
Finland	0.223	-0.009	-0.224	-0.454	-0.320	-0.075
France	-1.097**	-1.329**	-1.543**	-1.774***	-1.639***	-1.395**
Germany	0.576	0.344	0.129	-0.101	0.034	0.278
Italy	0.300	0.067	-0.147	-0.378	-0.243	0.001
United Kingdom	-0.780	-1.013*	-1.227**	-1.458***	-1.323	-1.079**
United States	-0.186	-0.418	-0.633	-0.863*	-0.728	-0.484

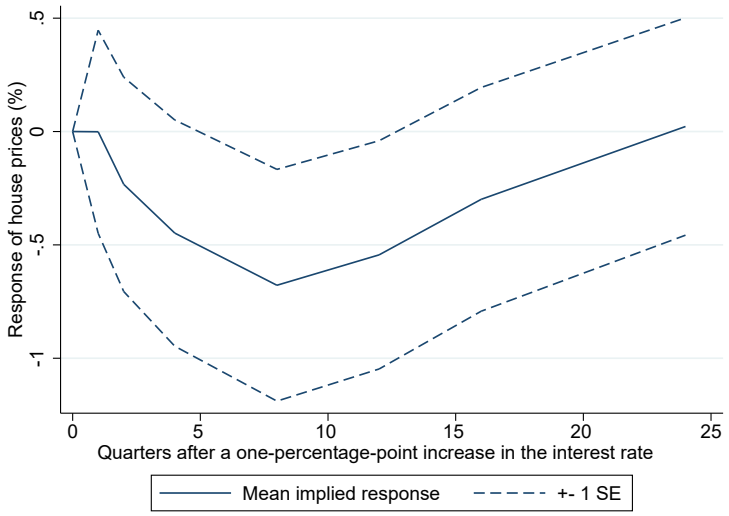
Notes: The values represent the percentage response of house prices to a one-percentage-point increase in the policy rate. They correspond to mean estimates conditional on best practice in the literature (see text for more details) and are computed based on fitted values from Bayesian model averaging (for example, by substituting "0" for the standard error, "1" for the inclusion of long-run interest rates, and so on). The second row shows results conditional mean values for specification characteristics. The estimates for individual countries are based on the baseline definition of best practice. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level; significance is approximate and based on standard errors obtained from frequentist model averaging.

sample means—of course, in the case of impulse responses constructed for individual countries we set the structural variables to the mean values corresponding to the individual countries.

The results are shown in Table 3 and Figure 8. While the main analysis in this section is based on the four-quarter horizon for ease of exposition, in order to compute the implied impulse response we need to run BMA analyses for each horizon separately. The corresponding analyses are not reported here, but in Table C3 in Appendix C we present the concise results of OLS estimates for each horizon. For each horizon the implied semi-elasticity is computed using the approach described in the previous two paragraphs. The first row of the table shows the baseline implied response, which is also depicted graphically in Figure 8. At no horizon is the impulse response significantly different from zero at the 5% level, but at the eight-quarter horizon it is marginally significant at the 32% level commonly used in vector autoregressions. The implied uncertainty in transmission is large, and the 68% confidence interval excludes -1.2, the mean maximum response of house prices uncorrected for publication bias. The mean maximum corrected semi-elasticity is -0.68, which suggests practically unimportant transmission of monetary policy to house prices—on average at least.

The second row of Table 3 presents the results of the same exercise with the exception of the preferred values for specification characteristics. While we prefer the inclusion of controls for liquidity and long-term interest rates, the preference is not universal among the most prominently published studies in the literature. So as a robustness check, we compute the implied impulse response without any preference on these variables (that is, we plug in sample means for all the variables reflecting specification characteristics). The resulting responses of house prices are substantially larger with the semi-elasticity reaching -1.4 after eight quarters. Still the response is not large enough to be of practical importance in taming housing bubbles. It follows that different specification of the VAR model can easily change the estimated response

Figure 8: Impulse response corrected for publication bias and misspecifications



Notes: The figure shows the mean impulse response reported in the literature and conditional on preferred aspects of data, methods, and publication. Based on the baseline exercise computed in Table 3.

of house prices by around one percentage point. In the remaining rows of Table 3 we compute impulse responses for several selected countries. Even the strongest semi-elasticity (-1.8 in France) is insufficient for plausible leaning against the wind when house prices inflation reaches double digits. The weakest semi-elasticity appears again in Germany (-0.1 after two years), which means that country-level characteristics can explain differences of up to 1.7 in the semi-elasticity. The differences explained by country and business-cycle characteristics can rise up to 3 if we select extreme values for these characteristics (not reported in the table). But even the impulse responses implied by the most extreme outliers in the values of these characteristics suggest semi-elasticities above -3.

5 Concluding Remarks

We collect 1,447 estimates of the reaction of house prices to a monetary policy shock at different horizons reported in 221 impulse responses from 31 studies. After correcting for publication bias and misspecifications, our results suggest that a one-percentage-point increase in the policy rate is on average associated with a maximum decrease of 0.7% in house prices after two years. The estimate has wide confidence intervals, which suggest that the transmission of monetary policy to house prices is uncertain and unstable in addition to being typically weak. Indeed, we find that transmission varies substantially across countries and time: it is stronger in countries with more developed mortgage markets and in the latter part of the business cycle. But even the most optimistic estimates for the periods and countries with characteristics conducive to more effective transmission imply semi-elasticities of less than 3 in the absolute value. So while leaning against the wind may help partly mitigate housing bubbles, the policy rate is a crude

instrument for such a task and one costly in terms of inflation and unemployment. Svensson (2017) compares the benefits and costs of leaning against the wind and comes to the conclusion that in most contexts costs outweight benefits by a large margin. Targeted macroprudential policy tools in the form of binding loan-to-income or debt-service-to-income ratios appear more likely to succeed in steering house prices, although empirical evidence on their effectiveness is still relatively thin (Poghosyan, 2020).

Three qualifications of our results are in order. First, in a way unusual but not unheard of in meta-analysis (Fabo *et al.*, 2021), we collect data from graphical results (impulse responses and the corresponding confidence intervals). Even though we do our best to codify the numerical values as precisely as possible, a random classical measurement error inevitably arises. In a regression of the estimated semi-elasticity on the corresponding standard error, therefore, the slope coefficient is biased downward due to attenuation bias. Because in our benchmark models the slope coefficient measures the strength of publication bias, many of our estimations are likely to underestimate the effects of the bias and hence produce conservative corrections. In fact, the problem with measurement error is more benign in the synthesis of graphical results than in the traditional synthesis of numerical results. The reason is that numerical results are rounded. Because different studies round differently, measurement error might not be random across studies. Bruns *et al.* (2019) show that rounding can create a false impression of publication bias (for example, the clustering of t-statistics at integers such as 2).

Second, the baseline meta-analysis models that we use come from or are inspired by medical research. In medical research, it is common to assume that the standard error is given to the researcher, often directly proportional to the number of subjects. That is, the standard error is exogenous and in the absence of publication bias there should be no correlation between estimates and standard errors. But in economics the computation of the standard error forms an important part of the exercise: in the VAR literature, for example, the confidence intervals can be constructed using different bootstrapping approaches, and different estimation techniques will generally yield different intervals. It follows that publication bias can also work via unintentional manipulation of the reported precision, not only the reported point estimate as is commonly assumed in meta-analysis. One solution is to use a function of the number of observations as an instrument for the standard error, but in the VAR literature the instrument is weak. We thus employ the new p-uniform* technique (van Aert & van Assen, 2021) developed in psychology, which uses the distribution of p-values and assumes nothing about the relationship between estimates and standard errors. As robustness checks we also use the techniques by Gerber & Malhotra (2008a), Simonsohn *et al.* (2014a), and Elliott *et al.* (2021) that too do not need the exogeneity assumption.

Third, in this meta-analysis we ignore the growing literature on the effects of unconventional monetary policy on house prices (see, for example, Rahal, 2016; Lenza & Slacalek, 2018; Rosenberg, 2019). While the short-term policy rate appears to have only limited influence on house prices, other tools of monetary policy (such as quantitative easing) might have played a more prominent role recently. Indeed, our results indicate that controlling for liquidity reduces

the reported effects of policy rates on house prices, which suggests that tools which primarily affect liquidity can be important. But the studies focusing on unconventional policy are quantitatively incomparable with the rest of our sample, and we believe they are best analyzed separately in a future research synthesis. The literature also lacks a thorough synthesis on the effects of macroprudential policies on house prices. As the body of relevant empirical research grows, conducting a meta-analysis will soon be possible in both realms.

References

van Aert, R. C. & M. van Assen (2021): "Correcting for publication bias in a meta-analysis with the p-uniform* method." *Working paper*, Tilburg University.

Amini, S. M. & C. F. Parmeter (2012): "Comparison of model averaging techniques: Assessing growth determinants." *Journal of Applied Econometrics* **27(5)**: pp. 870–876.

Andrews, I. & M. Kasy (2019): "Identification of and correction for publication bias." *American Economic Review* **109(8)**: pp. 2766–2794.

Ashenfelter, O. & M. Greenstone (2004): "Estimating the Value of a Statistical Life: The Importance of Omitted Variables and Publication Bias." *American Economic Review* **94(2)**: pp. 454–460.

Ashenfelter, O., C. Harmon, & H. Oosterbeek (1999): "A review of estimates of the schooling/earnings relationship, with tests for publication bias." *Labour Economics* **6(4)**: pp. 453–470.

Assenmacher-Wesche, K. & S. Gerlach (2010): "Monetary policy and financial imbalances: Facts and fiction." *Economic Policy* **25(63)**: pp. 437–482.

Bjørnland, H. C. & D. H. Jacobsen (2010): "The role of house prices in the monetary policy transmission mechanism in small open economies." *Journal of Financial Stability* **6(4)**: pp. 218–229.

Blanco-Perez, C. & A. Brodeur (2020): "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* **130(629)**: pp. 1226–1247.

Brodeur, A., N. Cook, & A. Heyes (2020): "Methods matter: P-hacking and publication bias in causal analysis in economics." *American Economic Review* **110(11)**: pp. 3634–60.

Brodeur, A., M. Lé, M. Sangnier, & Y. Zylberberg (2016): "Star wars: The empirics strike back." *American Economic Journal: Applied Economics* **8(1)**: pp. 1–32.

Bruns, S. B., I. Asanov, R. Bode, M. Dunger, C. Funk, S. M. Hassan, J. Hauschildt, D. Heinisch, K. Kempa, J. König *et al.* (2019): "Reporting errors and biases in published empirical findings: Evidence from innovation research." *Research Policy* **48(9)**: p. 103796.

Bruns, S. B. & J. P. A. Ioannidis (2016): "p-Curve and p-Hacking in Observational Research." *PloS ONE* **11(2)**: p. e0149144.

Calza, A., T. Monacelli, & L. Stracca (2013): "Housing finance and monetary policy." *Journal of the European Economic Association* **11(1)**: pp. 101–122.

Cameron, A. C., J. B. Gelbach, & D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics* **90(3)**: pp. 414–427.

Card, D., J. Kluve, & A. Weber (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* **16(3)**: pp. 894–931.

Card, D. & A. B. Krueger (1995): "Time-series minimum-wage studies: A meta-analysis." *The American Economic Review* **85(2)**: pp. 238–243.

Cattaneo, M. D., M. Jansson, & X. Ma (2020): "Simple local polynomial density estimators." *Journal of the American Statistical Association* **115(531)**: pp. 1449–1455.

Christensen, G. & E. Miguel (2018): "Transparency, reproducibility, and the credibility of economics research." *Journal of Economic Literature* **56(3)**: pp. 920–980.

Del Negro, M. & C. Otrok (2007): "99 Luftballons: Monetary Policy and the House Price Boom Across US States." *Journal of Monetary Economics* **54(7)**: pp. 1962–1985.

DellaVigna, S., D. Pope, & E. Vivalt (2019): "Predict science to improve science." *Science* **366(6464)**: pp. 428–429.

Egger, M., G. D. Smith, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *Journal of Economic Surveys* **315(7109)**: p. 629–634.

Eicher, T. S., C. Papageorgiou, & A. E. Raftery (2011): "Default priors and predictive performance in bayesian model averaging, with application to growth determinants." *Journal of Applied Econometrics* **26(1)**: pp. 30–55.

Elliott, G., N. Kudrin, & K. Wuthrich (2021): "Detecting p-hacking." *Econometrica* **(forthcoming)**.

Fabo, B., M. Jancokova, E. Kempf, & L. Pastor (2021): "Fifty Shades of QE: Comparing Findings of Central Bankers and Academics." *Journal of Monetary Economics* **(forthcoming)**.

Fernandez, C., E. Ley, & M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions." *Journal of Applied Econometrics* **16(5)**: pp. 563–576.

Fratantoni, M. & S. Schuh (2003): "Monetary policy, housing, and heterogeneous regional markets." *Journal of Money, Credit and Banking* **35(4)**: pp. 557–589.

Furukawa, C. (2021): "Publication Bias under Aggregation Frictions: Theory, Evidence, and a New Correction Method." *Working paper*, MIT.

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." In "Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown," pp. 158–165. Institute of Mathematical Statistics.

Gerber, A. & N. Malhotra (2008a): "Do statistical reporting standards affect what is published? publication bias in two leading political science journals." *Quarterly Journal of Political Science* **3(3)**: pp. 313–326.

Gerber, A. S. & N. Malhotra (2008b): "Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?" *Sociological Methods & Research* **37(1)**: pp. 3–30.

Grimes, A. & A. Aitken (2010): "Housing supply, land costs and price adjustment." *Real Estate Economics* **38(2)**: pp. 325–353.

Hansen, B. E. (2007): "Least squares model averaging." *Econometrica* **75(4)**: p. 1175–1189.

Havranek, T. (2015): "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting." *Journal of the European Economic Association* **13(6)**: pp. 1180–1204.

Havranek, T., T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, & R. C. M. van Aert (2020): "Reporting Guidelines for Meta-Analysis in Economics." *Journal of Economic Surveys* **34(3)**: pp. 469–475.

Iacoviello, M. & R. Minetti (2003): "Financial liberalization and the sensitivity of house prices to monetary policy: Theory and evidence." *The Manchester School* **71(1)**: pp. 20–34.

Iacoviello, M. & S. Neri (2010): "Housing market spillovers: Evidence from an estimated dsge model." *American Economic Journal: Macroeconomics* **2(2)**: pp. 125–164.

Imai, T., T. A. Rutter, & C. F. Camerer (2021): "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets." *Economic Journal* **131(636)**: pp. 1788–1814.

Ioannidis, J. P., T. Stanley, & H. Doucouliagos (2017): "The power of bias in economics research." *Economic Journal* **127(605)**: pp. 236–265.

Lenza, M. & J. Slacalek (2018): "How does monetary policy affect income and wealth inequality? Evidence from quantitative easing in the euro area." *Working Paper Series 2190*, European Central Bank.

Ley, E. & M. F. Steel (2009): "On the effect of prior assumptions in bayesian model averaging with applications to growth regression." *Applied Econometrics* **24**: pp. 651–674.

Magnus, J. R., O. Powell, & P. Prufer (2010): "A comparison of two model averaging techniques with an application to growth empirics." *Journal of Econometrics* **154(2)**: pp. 139–153.

McCloskey, D. & S. T. Ziliak (2019): "What Quantitative Methods Should We Teach to Graduate Students? A Comment on Swann's "Is Precise Econometrics an Illusion?"." *The Journal of Economic Education* **50(4)**: pp. 356–361.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, & D. T. Mellor (2018): "The preregistration revolution." *Proceedings of the National Academy of Sciences* **115(11)**: pp. 2600–2606.

Olken, B. A. (2015): "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* **29(3)**: pp. 61–80.

Paciorek, A. (2013): "Supply constraints and housing market dynamics." *Journal of Urban Economics* **77**: pp. 11–26.

Poghosyan, T. (2020): "How effective is macroprudential policy? Evidence from lending restriction measures in EU countries." *Journal of Housing Economics* **49(C)**.

Powell, T. & D. Wessel (2021): "Why is the New Zealand government telling its central bank to focus on rising house prices?" Brookings, Hutchins Center on Fiscal and Monetary Policy, April 2, 2021.

Rachel, L. & T. D. Smith (2017): "Are Low Real Interest Rates Here to Stay?" *International Journal of Central Banking* **13(3)**: pp. 1–42.

Raftery, A. E. (1995): "Bayesian model selection in social research." *Sociological Methodology* pp. 111–163.

Raftery, A. E., D. Madigan, & J. A. Hoeting (1997): "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* **92(437)**: pp. 179–191.

Rahal, C. (2016): "Housing markets and unconventional monetary policy." *Journal of Housing Economics* **32**: pp. 67–80.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, & M. D. Webb (2019): "Fast and Wild: Bootstrap Inference in Stata Using Boottest." *The Stata Journal* **19(1)**: pp. 4–60.

Rosenberg, S. (2019): "The effects of conventional and unconventional monetary policy on house prices in the Scandinavian countries." *Journal of Housing Economics* **46(C)**.

Rusnak, M., T. Havranek, & R. Horvath (2013): "How to solve the price puzzle? a meta-analysis." *Journal of Money, Credit and Banking* **45(1)**: pp. 37–70.

Siegfried, J. J. (2012): "Minutes of the Meeting of the Executive Committee: Chicago, IL, January 5, 2012." *American Economic Review* **102(3)**: pp. 645–652.

Simonsohn, U., L. D. Nelson, & J. P. Simmons (2014a): "P-curve: A key to the file-drawer." *Journal of Experimental Psychology: General* **143(2)**: pp. 534–547.

Simonsohn, U., L. D. Nelson, & J. P. Simmons (2014b): "P-curve and effect size: Correcting for publication bias using only significant results." *Perspectives on Psychological Science* **9(6)**: pp. 666–681.

Sims, C. A. (1980): "Macroeconomics and reality." *Econometrica* **48(1)**: pp. 1–48.

Stanley, T. D. (2001): "Wheat from chaff: Meta-analysis as quantitative literature review." *Journal of Economic Perspectives* **15(3)**: pp. 131–150.

Stanley, T. D. (2005): "Beyond publication bias." *Journal of Economic Surveys* **19(3)**: pp. 309–345.

Stanley, T. D. (2008): "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics* **70(1)**: pp. 103–127.

Steel, M. F. (2020): "Model averaging and its use in economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Strømland, E. (2019): "Preregistration and reproducibility." *Journal of Economic Psychology* **75(PA)**.

Svensson, L. E. (2014): "Inflation Targeting and "Leaning against the Wind"." *International Journal of Central Banking* **10(2)**: pp. 103–114.

Svensson, L. E. (2017): "Cost-benefit analysis of leaning against the wind." *Journal of Monetary Economics* **90(C)**: pp. 193–213.

Taylor, J. B. (2007): "Housing and monetary policy." *NBER Working Paper No. 13682*, National Bureau of Economic Research.

Uhlig, H. (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure." *Journal of Monetary Economics* **52(2)**: pp. 381–419.

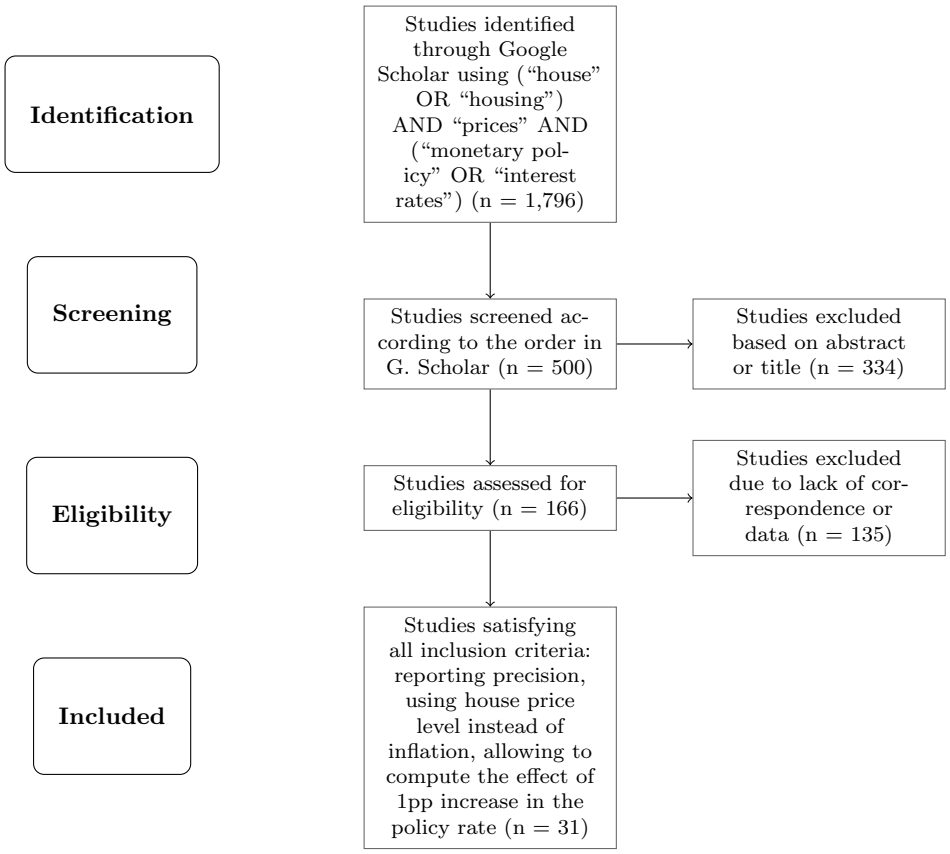
Uhlig, H. (2012): "Economics and reality." *Journal of Macroeconomics* **34(1)**: pp. 29–41.

Williams, J. C. (2016): "Measuring the effects of monetary policy on house prices and the economy." *BIS Papers No. 88*, Bank for International Settlements.

Zeugner, S. & M. Feldkircher (2015): "Bayesian model averaging employing fixed and flexible priors: The bms package for r." *Journal of Statistical Software* **68(4)**: pp. 1–37.

A Details of Literature Search

Figure A1: PRISMA flow diagram



Notes: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) is an evidence-based set of items for reporting in systematic reviews and meta-analyses. More details on PRISMA and reporting standard of meta-analysis in general are provided by Havranek *et al.* (2020).

B Extensions of Publication Bias Models (for Online Publication)

B.1 Caliper Test

As an extension to the previously reported tests of publication bias we apply the caliper test as proposed in Gerber & Malhotra (2008a) and Gerber & Malhotra (2008b) and recently implemented by Bruns *et al.* (2019). The caliper test is based on the analysis of discontinuities in the reported t-statistics: if no selective reporting is present, there should be no discontinuities around the conventional significance thresholds. In other words the number of t-statistics reported in the literature just above the threshold ("over caliper") should not be statistically different from the number of reported t-statistics just below the threshold ("under caliper"). The test does not allow us to compute the true effect beyond bias but serves as an indicator of whether publication selection appears in the literature, thus providing us with a robustness check of the previous results. The results are presented in Table B1. Primarily we examine the significance threshold corresponding to the 68% confidence interval: although the threshold is usually much stricter in the empirical literature featuring point estimates, in the case of VAR models and impulse response functions the 68% confidence interval is the most frequently reported (almost 70% of our estimates use it), so we suspect that publication selection could be related to this threshold. We use caliper sizes of 0.1, 0.3, and 0.5. The results show that publication selection is present at the horizons of eight quarters and one quarter. If we test the parameter against the value of 0.4 (i.e., a 60:40 distribution around the thresholds, instead of 50:50, as reasoned in Bruns *et al.* 2019), then evidence of publication selection is also present at the horizon of four quarters and when all the horizons are tested together. This is broadly in line with our previous results on publication selection.

Table B1: Results of the caliper test

Caliper size	All horizons	1 quarter	2 quarters	4 quarters	8 quarters	12 quarters	16 quarters
0.1	*0.521*	**0.722**	0.500	0.444	0.429	0.467	0.471
(95% LCI)	*(0.436)*	**(0.533)**	(0.289)	(0.118)	(0.035)	(0.231)	(0.252)
0.3	*0.527*	**0.625**	0.512	*0.556*	**0.625**	0.477	0.434
(95% LCI)	*(0.482)*	**(0.507)**	(0.379)	*(0.430)*	**(0.516)**	(0.349)	(0.319)
0.5	*0.509*	**0.613**	0.452	*0.560*	**0.595**	0.486	0.422
(95% LCI)	*(0.474)*	**(0.521)**	(0.354)	*(0.464)*	**(0.506)**	(0.387)	(0.331)

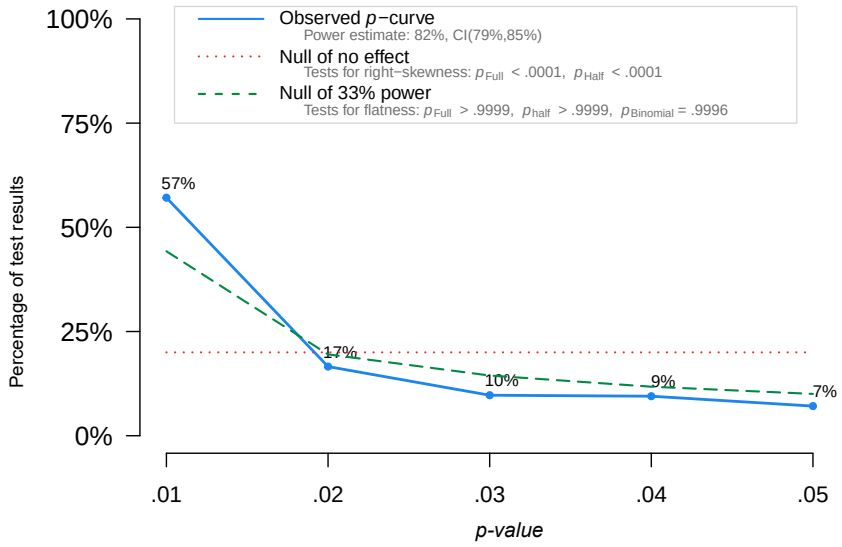
Notes: The table shows the results of the caliper test for three caliper sizes 0.1, 0.3, and 0.5. The reported numbers represent the share of observations in the narrow interval that are above the significance threshold. LCI = lower bound of the confidence interval. The test parameter is the following: $C = \frac{n_{oc}}{n_{oc}+n_{uc}}$, where n_{oc} and n_{uc} stand for the number of observations with t-statistics in the interval above the threshold ("over caliper") and below the threshold ("under caliper"). For the significance threshold we use the criterion of one standard error above the estimate (commonly used in the VAR literature). The one-sided hypothesis $H_0 : C \leq 0.5$ is tested against $H_1 : C > 0.5$. 95% lower confidence intervals for the test parameters are reported in parenthesis. Significant caliper test results when testing $H_0 : C \leq 0.5$ are shown in bold; significant caliper test results when testing $H_0 : C \leq 0.4$ are shown in italics.

B.2 Tests Based on the Distribution of p-values

B.2.1 p-curve

Now we look at the distribution of p-values. First, we employ the p-curve method, which is primarily intended to test the null hypothesis that the literature has no evidential value (that is, no effect of monetary policy on house prices beyond publication bias). The technique was developed by Simonsohn *et al.* (2014a) and Simonsohn *et al.* (2014b). Based on Figure B1, we obtain evidence for evidential value, which is consistent with a right-skewed distribution, while a left-skewed distribution would suggest p-hacking. In addition to contrast to the common p-curve we also plot the whole distribution of p-values (not only those significant up to the 5% significance level) in Figure B2 to see whether there are distinct jumps at different thresholds associated with conventional statistical significance. There is no clear evidence for such jumps.

Figure B1: p-curve results

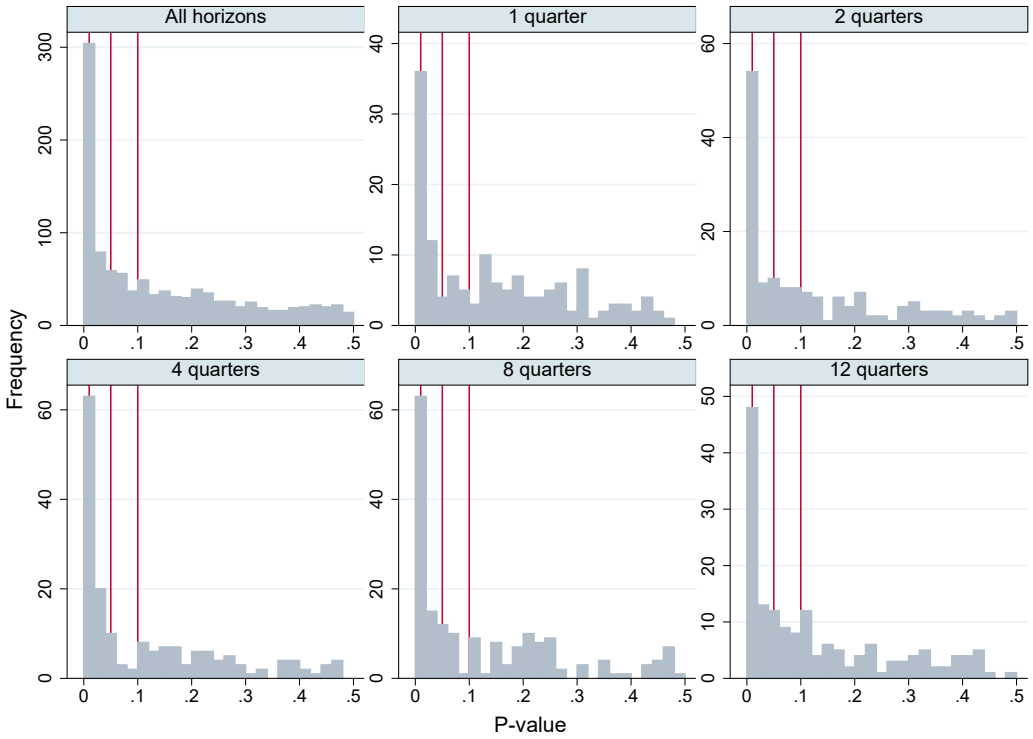


Note: The observed p−curve includes 422 statistically significant (p < .05) results, of which 331 are p < .025.
There were 1025 additional results entered but excluded from p−curve because they were p > .05.

B.2.2 Tests introduced by Elliott *et al.* (2021)

Elliott *et al.* (2021) analyze p-hacking based on the distribution of p-values and introduce novel testable restrictions. They show that the p-curve (distribution of p-values across studies) is i) non-increasing and continuous in the absence of p-hacking, ii) completely monotone, with upper bounds on p-curve. In their empirical application they use binomial, Fisher's, and density discontinuity tests, as already used before in Simonsohn *et al.* (2014a) and Cattaneo *et al.* (2020). Besides that, they also develop new, more powerful tests: a histogram-based test for non-increasingness, a histogram-based test for 2-monotonicity and bounds, and least concave majorant (LCM) test based on concavity of the CDF of p-values. The results of these tests are available in Table B2. All of the tests have null hypothesis of no p-hacking. While with

Figure B2: Distribution of p-values



less powerful tests (binomial and Fisher) we do not reject the null of no p-hacking, we can reject it with the test for non-increasingness (CS1), 2-monotonicity (CS2B) and also density discontinuity test at horizons between 2 and 12 quarters in all cases. As in the main body of the paper, we run the tests at a threshold of t=1, instead of 1.96, as this is the most common threshold in impulse responses of VAR models.

Table B2: Tests used by Elliott *et al.* (2021)

Test	All	1 Quarter	2 Quarters	4 Quarters	8 Quarters	12 Quarters	16 Quarters
Binomial	0.990	0.702	0.500	0.837	1.000	0.820	0.500
Fisher	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	**0.000**	0.161	**0.000**	**0.000**	**0.001**	**0.015**	0.671
CS1	**0.069**	0.891	**0.053**	**0.000**	**0.019**	**0.000**	0.194
CS2B	**0.000**	**0.000**	**0.008**	**0.000**	**0.000**	**0.000**	**0.004**
LCM	1.000	0.987	0.999	0.996	0.691	0.997	0.994
N near t=1	178	32	21	26	35	19	23
N	1054	142	156	174	182	164	146

Notes: CS1 is the test for non-increasingness. CS2B is the test for K-monotonicity. LCM is the test based on the concavity of the CDF of p-values. Values in bold incicate rejections of the hypothesis of no p-hacking.

C Summary Statistics and Extensions of Heterogeneity Models (for Online Publication)

Table C1: Description and summary statistics of regression variables

Label	Description	Mean	SD
Estimate	The reported effect of a one-percentage-point increase in the interest rate on house prices (after four quarters in %).	-0.849	1.126
Standard Error	The reported or implied standard error of the estimate.	0.765	0.878
Data characteristics			
Monthly	= 1 if the data were collected at the monthly frequency (reference category: quarterly data).	0.096	0.295
Panel	= 1 if panel data were used (ref. cat.: time series).	0.208	0.406
Length	The logarithm of the length of the data sample used in the primary study (in years).	3.102	0.283
Midpoint	The logarithm of the mean year of the data used in the study (normalized to the earliest mean year in our sample).	2.862	0.500
Specification characteristics			
GDP Defl.	= 1 if GDP deflator is included in the VAR model instead of CPI.	0.075	0.263
Foreign IR	= 1 if a foreign interest rate is included.	0.028	0.164
Credit	= 1 if credit is included.	0.261	0.439
Consumption	= 1 if consumption is included.	0.294	0.456
Res. Invest	= 1 if a measure of residential investment is included.	0.185	0.388
Money Supply	= 1 if a measure of the money supply is included.	0.191	0.393
Exch. Rate	= 1 if the exchange rate is included	0.233	0.423
Long-run IR	= 1 if the long-run interest rate (in addition to the short-run interest rate) is included.	0.168	0.374
Real HP	= 1 if real instead of nominal house prices are used.	0.950	0.218
Lags	The number of lags (in quarters) included in the model.	3.256	1.265
Time Trend	= 1 if the study uses detrended data or a time trend is added to the regression.	0.395	0.489
Estimation characteristics			
BVAR	= 1 if a Bayesian VAR model is employed in the primary study.	0.095	0.294
Sign Restr. HP	= 1 if sign restrictions are used in the VAR model and are imposed on the house price variable (ref. cat.: Cholesky decomposition).	0.052	0.222
Sign Restr. Other	= 1 if sign restrictions are used in the VAR model but are not imposed on the house price variable (ref. cat.: Cholesky decomposition).	0.029	0.169
Nonrecursive	= 1 if another nonrecursive identification is used in the VAR model (ref. cat.: Cholesky decomposition).	0.124	0.329
Publication characteristics			
Citations	The logarithm of the number of citations of the study per year since its first appearance in Google Scholar.	1.939	0.657

Continued on next page

Table C1: Description and summary statistics of regression variables (Continued)

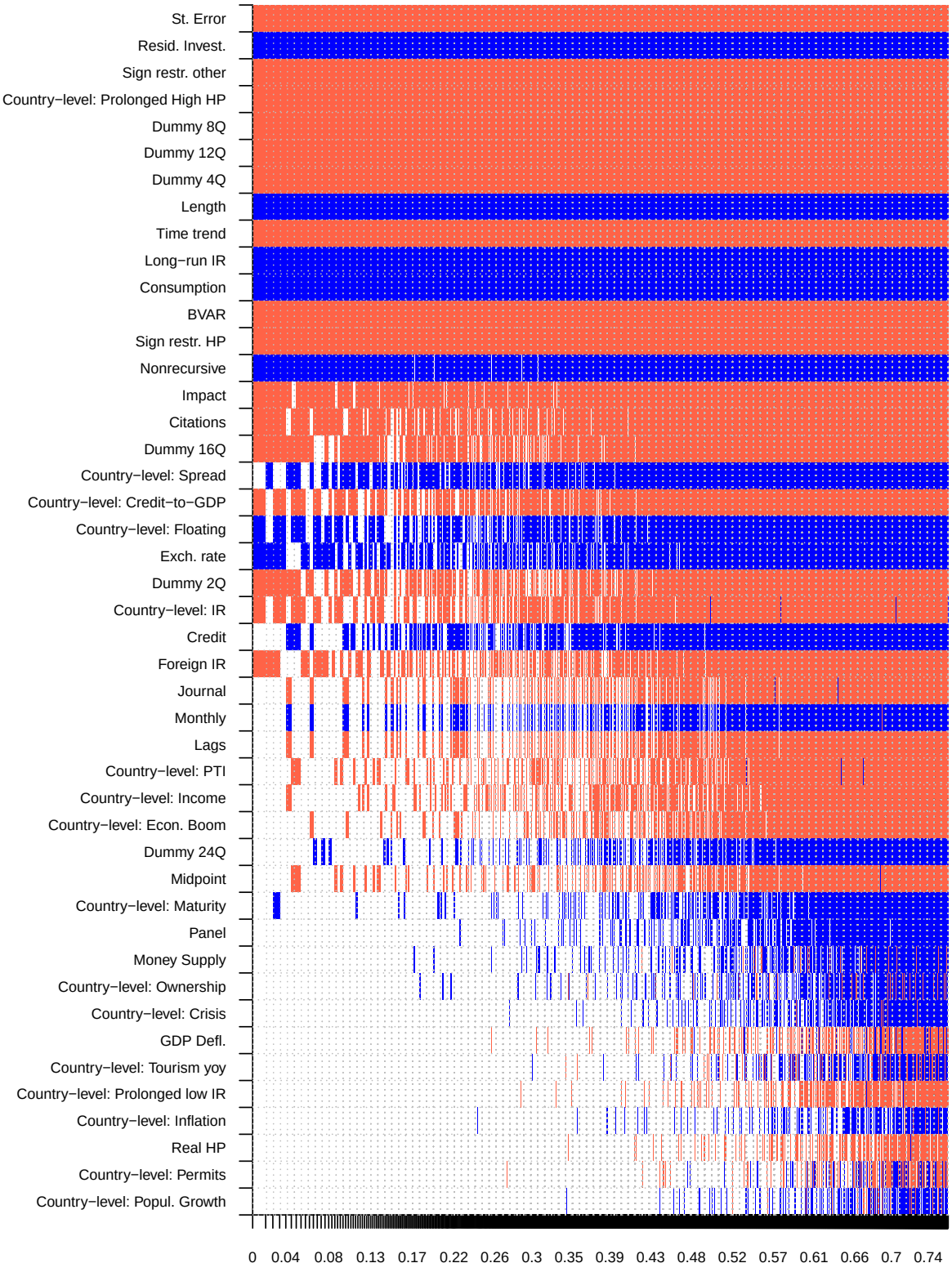
Label	Description	Mean	SD
Impact	The recursive discounted RePEc impact factor of the outlet.	0.478	0.504
Journal	= 1 if the study is published in a peer-reviewed journal.	0.366	0.482
Structural heterogeneity (computed for the period and country for which the VAR is estimated)			
Crisis	The number of years (out of those used in the time span of the primary study) during which a banking crisis occurred.	3.356	2.679
IR	The average three-month interest rate, OECD.	7.225	2.495
Prolonged Low IR	The number of consecutive years (out of those used in the time span of the primary study) during which the short-run interest rate was below its long-run average.	8.578	4.744
Spread	The average difference between short-term and long-term interest rates.	0.660	0.495
Floating	The share of loans with a floating interest rate.	50.958	27.445
Tourism YoY	The growth rate of the number of arrivals per capita.	3.463	5.032
Income	Average disposable income per household per capita in US dollars, OECD.	9.753	0.362
Inflation	Average consumer price inflation, OECD.	4.266	2.370
Credit-to-GDP	The credit-to-GDP ratio, BIS.	124.785	33.476
Popul. Growth	Average annual population growth, World Bank.	0.608	0.400
PTI	The standardized price-to-income ratio for the housing market.	94.112	8.895
Prolonged High HP	The number of periods (out of those used in the time span of the primary study) with above-average house price growth.	12.180	4.662
Permits	The number of building permits issued in comparison to its long-run average.	101.394	21.977
Maturity	The average maturity of mortgage loans.	3.079	0.226
Ownership	The share of home ownership.	61.017	9.046
Econ. Boom	The number of periods (out of those used in the time span of the primary study) with a positive output gap.	5.484	3.367

Table C2: Results of frequentist model averaging

Category	Variable	Coef.	Std. Er.	p-value
Publication bias	**SE**	**-1.555**	**0.183**	**0.000**
Data characteristics	Monthly	0.185	0.552	0.737
	Panel	-0.026	0.284	0.927
	Length	0.887	0.711	0.212
	Midpoint	-0.024	0.287	0.934
Specification characteristics	GDP Defl.	0.075	0.303	0.805
	Foreign IR	-0.108	0.370	0.770
	Credit	0.275	0.222	0.215
	Consumption	0.249	0.293	0.395
	Resid. Invest.	0.415	0.390	0.287
	Money Supply	*0.473*	*0.352*	*0.179*
	Exch. rate	0.230	0.270	0.393
	Long-run IR	*0.495*	*0.322*	*0.125*
	Real HP	0.191	0.347	0.582
	Lags	0.030	0.085	0.726
	Time trend	-0.211	0.218	0.333
Estimation characteristics	BVAR	-0.727	0.627	0.246
	Sign restr. HP	*-0.685*	*0.484*	*0.157*
	Sign restr. other	-1.232	0.860	0.152
	Nonrecursive	0.102	0.387	0.792
Publication characteristics	Citations	-0.023	0.242	0.925
	Impact	-0.235	0.243	0.332
	Journal	-0.116	0.266	0.663
Structural heterogeneity	Country-level: Crisis	0.032	0.042	0.450
	Country-level: IR	-0.166	0.113	0.140
	Country-level: Prolonged low IR	-0.023	0.031	0.455
	Country-level: Spread	0.238	0.218	0.274
	Country-level: Floating	0.005	0.004	0.284
	Country-level: Tourism yoy	-0.009	0.015	0.557
	Country-level: Income	0.448	0.687	0.515
	Country-level: Inflation	0.068	0.069	0.325
	Country-level: Credit-to-GDP	**-0.013**	**0.007**	**0.077**
	Country-level: Popul. Growth	0.513	0.543	0.345
	Country-level: PTI	-0.030	0.019	0.127
	Country-level: Prolonged High HP	**-0.069**	**0.041**	**0.093**
	Country-level: Permits	0.005	0.007	0.451
	Country-level: Maturity	-0.248	0.474	0.601
	Country-level: Ownership	-0.017	0.019	0.368
	Country-level: Econ. Boom	-0.025	0.035	0.472
Observations	209			

Notes: The frequentist model averaging (FMA) exercise employs Mallow's weights (Hansen, 2007) and the orthogonalization of the covariate space suggested by Amini & Parmeter (2012). Variables significant at the 10% level are shown in bold; variables that were important in BMA and have a p-value lower than 0.2 are indicated in italics.

Figure C1: Model inclusion in BMA with estimates for all horizons



Notes: Columns denote individual models; the variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities; the 10,000 best models are shown. To ensure convergence we employ 3 million iterations and 1 million burn-ins. Blue color (darker in grayscale) = the variable is included and the estimated sign is positive, i.e., the transmission is weaker. Red color (lighter in grayscale) = the variable is included and the estimated sign is negative, i.e., the transmission is stronger. No color = the variable is not included in the model. A detailed description of all the variables is available in Table C1.

Table C3: A robustness check using ordinary least squares

Category	Variable	1Q	2Q	4Q	8Q	12Q	16Q
Publication bias	SE	-0.805***	-1.215***	-1.615***	-1.553***	-0.665***	-0.365***
		(0.204)	(0.294)	(0.278)	(0.168)	(0.230)	(0.122)
Data characteristics	Panel	-0.314***	-0.228	0.0151	0.166	0.439	0.752**
		(0.0822)	(0.141)	(0.186)	(0.268)	(0.286)	(0.294)
	Length	-0.386	0.874	1.170*	1.831***	0.644*	0.171
		(0.371)	(0.533)	(0.631)	(0.641)	(0.387)	(0.294)
Specification characteristics	GDP Defl	0.395***	0.638**	0.0965	-0.264	-0.244	-0.212
		(0.150)	(0.313)	(0.197)	(0.206)	(0.194)	(0.212)
	Credit	-0.0615	0.126*	0.299*	0.532***	0.0426	0.101
		(0.048)	(0.069)	(0.166)	(0.193)	(0.086)	(0.108)
	Consumption	0.115	0.295***	0.381**	0.514*	0.799***	0.811**
		(0.072)	(0.072)	(0.176)	(0.295)	(0.307)	(0.364)
	Resid. invest.	0.569***	0.610***	0.516**	0.494	0.918***	0.973***
		(0.176)	(0.188)	(0.223)	(0.324)	(0.241)	(0.272)
	Money Supply	-0.229**	-0.0266	0.527***	0.747***	0.0456	-0.102
		(0.117)	(0.152)	(0.191)	(0.221)	(0.351)	(0.377)
	Exchange rate	-0.053	0.175***	0.350***	0.529***	0.741***	0.741***
		(0.075)	(0.056)	(0.049)	(0.133)	(0.105)	(0.105)
	Long-run IR	0.0225	0.266**	0.577***	0.762***	0.574***	0.325***
		(0.070)	(0.107)	(0.156)	(0.212)	(0.0582)	(0.0343)
	Real HP	-0.443**	-0.230	0.138	0.279**	-0.426	-0.404
		(0.202)	(0.196)	(0.134)	(0.136)	(0.352)	(0.429)
	Lags	-0.084**	-0.047	0.049	-0.029	-0.098**	-0.184***
		(0.041)	(0.047)	(0.031)	(0.058)	(0.040)	(0.044)
	Time trend	-0.119	-0.308**	-0.315***	-0.637***	-0.531***	-0.500***
		(0.114)	(0.122)	(0.110)	(0.178)	(0.142)	(0.159)
Estimation characteristics	BVAR	-0.610*	-1.054**	-0.813*	-0.724	-1.488***	-1.038***
		(0.332)	(0.478)	(0.442)	(0.460)	(0.463)	(0.352)
	Sign restr. HP	-0.692***	-1.021***	-0.830*	-0.208	-0.313	0.275
		(0.185)	(0.369)	(0.428)	(0.471)	(0.476)	(0.392)
	Sign restr. other	0.232	-0.629*	-1.567***	-2.613***	-2.948***	-2.787***
		(0.167)	(0.330)	(0.303)	(0.728)	(0.480)	(0.500)
	Nonrecursive	0.704***	0.761**	0.305	0.164	1.307***	1.331***
		(0.190)	(0.306)	(0.293)	(0.309)	(0.398)	(0.446)
Publication characteristics	Citations	-0.159*	-0.153	-0.132	-0.164	-0.582***	-0.543***
		(0.095)	(0.183)	(0.225)	(0.215)	(0.153)	(0.203)
	Impact	-0.082	-0.252	-0.289*	-0.480	-0.660***	-0.639***
		(0.092)	(0.155)	(0.165)	(0.325)	(0.143)	(0.163)
Structural heterogeneity	Country-level: IR	-0.142**	-0.259**	-0.233**	-0.261***	-0.237***	-0.195***
		(0.070)	(0.111)	(0.110)	(0.066)	(0.029)	(0.017)
	Country-level: Spread	-0.255***	-0.108	0.287*	0.436***	0.419***	0.236*
		(0.097)	(0.202)	(0.157)	(0.122)	(0.124)	(0.142)
	Country-level: Floating	0.004	0.004	0.004	0.008***	0.008**	0.009**
		(0.005)	(0.005)	(0.006)	(0.003)	(0.003)	(0.004)
	Country-level: Tourism YoY	-0.023*	-0.030**	-0.007	0.004	0.005	-0.009
		(0.012)	(0.015)	(0.017)	(0.014)	(0.009)	(0.011)
	Country-level: Inflation	0.056	0.144***	0.096**	0.032	0.051	0.000
		(0.044)	(0.041)	(0.039)	(0.041)	(0.038)	(0.033)
	Country-level: Credit-to-GDP	-0.001	-0.012	-0.016*	-0.017***	-0.016***	-0.011***
		(0.005)	(0.007)	(0.008)	(0.005)	(0.000)	(0.002)
	Country-level: Maturity	0.776***	0.0632	-0.329	-0.259	0.808***	1.181***
		(0.240)	(0.711)	(0.717)	(0.611)	(0.237)	(0.391)
	Country-level: PTI	-0.019*	-0.024	-0.039***	-0.046***	-0.016**	-0.011
		(0.010)	(0.023)	(0.014)	(0.005)	(0.007)	(0.008)
	Country-level: Long High HP	-0.012	-0.083**	-0.090**	-0.105**	-0.115***	-0.090***
		(0.009)	(0.033)	(0.039)	(0.041)	(0.018)	(0.020)
	Observations	196	199	209	209	204	203

Notes: Standard errors, clustered at the study level, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. For ease of exposition, variables which are not significant at any horizon are excluded from the table.