DISCUSSION PAPER SERIES

DP16032

Publication and Identification Biases in Measuring the Intertemporal Substitution of Labor Supply

Tomas Havranek, Roman Horvath and Ali Elminejad

LABOUR ECONOMICS

CEPR

ISSN 0265-8003

Publication and Identification Biases in Measuring the Intertemporal Substitution of Labor Supply

Tomas Havranek, Roman Horvath and Ali Elminejad

Discussion Paper DP16032
Published 12 April 2021
Submitted 09 April 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Labour Economics

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Tomas Havranek, Roman Horvath and Ali Elminejad

Publication and Identification Biases in Measuring the Intertemporal Substitution of Labor Supply

Abstract

The intertemporal substitution (Frisch) elasticity of labor supply governs the predictions of real business cycle models and models of taxation. We show that, for the extensive margin elasticity, two biases conspire to systematically produce large positive estimates when the elasticity is in fact zero. Among 723 estimates in 36 studies, the mean reported elasticity is 0.5. One half of that number is due to publication bias: larger estimates are reported preferentially. The other half is due to identification bias: studies with less exogenous time variation in wages report larger elasticities. Net of the biases, the literature implies a zero mean elasticity and, with 95% confidence, is inconsistent with calibrations above 0.25. To derive these results we collect 23 variables that reflect the context in which the elasticity was obtained, use nonlinear techniques to correct for publication bias, and employ Bayesian and frequentist model averaging to address model uncertainty.

JEL Classification: C83, E24, J2

Keywords: Frisch elasticity, Labor Supply, extensive margin, meta-analysis, Publication bias, Bayesian model averaging

Tomas Havranek - tomas.havranek@ies-prague.org
Charles University, Prague and CEPR

Roman Horvath - roman.horvath@gmail.com
Charles University, Prague

Ali Elminejad - m.ali.elminejad@gmail.com
Charles University

Publication and Identification Biases in Measuring the Intertemporal Substitution of Labor Supply*

Ali Elminejad[a], Tomas Havranek[a,b], and Roman Horvath[a]

[a]Charles University, Prague

[b]CEPR

April 9, 2021

Abstract

The intertemporal substitution (Frisch) elasticity of labor supply governs the predictions of real business cycle models and models of taxation. We show that, for the extensive margin elasticity, two biases conspire to systematically produce large positive estimates when the elasticity is in fact zero. Among 723 estimates in 36 studies, the mean reported elasticity is 0.5. One half of that number is due to publication bias: larger estimates are reported preferentially. The other half is due to identification bias: studies with less exogenous time variation in wages report larger elasticities. Net of the biases, the literature implies a zero mean elasticity and, with 95% confidence, is inconsistent with calibrations above 0.25. To derive these results we collect 23 variables that reflect the context in which the elasticity was obtained, use nonlinear techniques to correct for publication bias, and employ Bayesian and frequentist model averaging to address model uncertainty.

Keywords: Frisch elasticity, labor supply, extensive margin, meta-analysis, publication bias, Bayesian model averaging

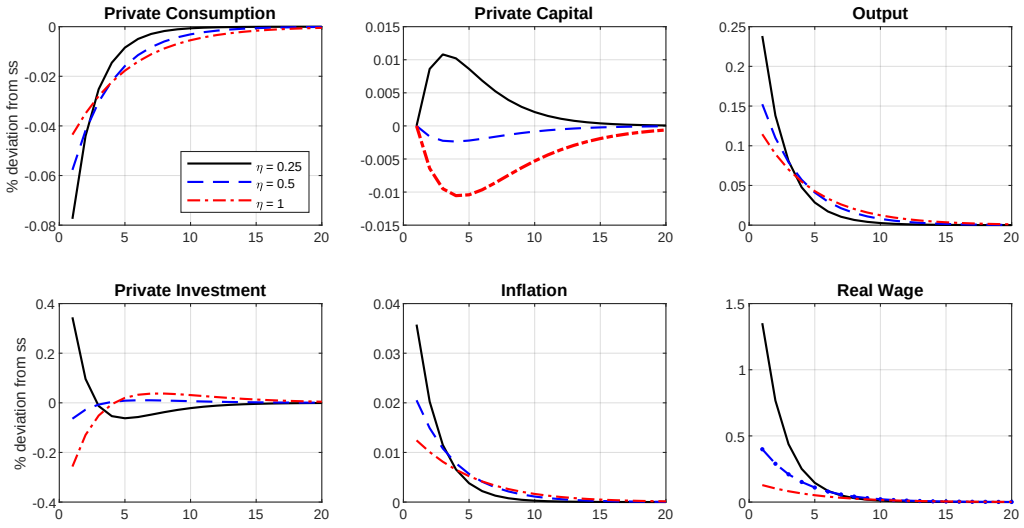
JEL Codes: C83, E24, J21

*An online appendix with data and code is available at `meta-analysis.cz/frisch`. Corresponding author: Roman Horvath, `roman.horvath@fsv.cuni.cz`.

1 Introduction

The Frisch elasticity of labor supply, the change in hours worked in response to changes in anticipated wages while keeping the marginal utility of wealth unchanged, plays the star part in answering central macroeconomic questions. How does labor supply react to technological shocks over the business cycle? How does a temporary tax increase affect the economy? And in general, what are the effects of fiscal policy? In Figure 1 we use the canonical New Keynesian model of Galí (2015) to illustrate the importance of the Frisch elasticity in modeling fiscal shocks, in this case a one-percentage-point increase in government spending. With different values of the elasticity we obtain very different stories in terms of the implied trajectory of private capital and private investment, but also total output and inflation. The Frisch elasticity clearly matters.

Figure 1: The Frisch elasticity drives the modeled impact of fiscal policy



Notes: The figure shows impulse responses to a one-percentage-point increase in government spending. The horizontal axis depicts quarters after the increase, the vertical axis depicts the percentage deviation from the steady state (ss). We use the standard model of Galí (2015) and change the value of the Frisch elasticity (η) while leaving all the other parameters calibrated at the values used by Galí (2015).

For calibrations, the authors of structural models have increasingly relied on the entire corpus of microeconomic empirical literature instead of cherry-picking one or two preferred results from the literature. A prominent example is the life-cycle model of the Congressional Budget Office (CBO), which relies on a careful survey of microeconomic evidence to calibrate the elasticity in the range 0.27–0.53 with a central estimate of 0.4 (Whalen & Reichling, 2017). The focus on the

entire literature is laudable, but in this paper we show that such a calibration proves misleading. The mean reported estimate is a poor and systematically biased reflection of the underlying quantity, an observation that extends beyond the Frisch elasticity and the CBO's model. While the CBO's calibration is not based on a formal meta-analysis, it matches our data remarkably well: the mean microeconomic estimate in our dataset is 0.41 (the overall mean, including macro estimates, is 0.49), and 56% of micro estimates lie between 0.27 and 0.53. Nevertheless, these summary statistics of the literature are heavily distorted by publication bias and endogeneity in some studies. Conditional on the absence of publication bias and the presence of arguably exogenous time variation in wages, the literature is consistent with zero Frisch elasticity at the extensive margin. While in the meta-analysis we do not include estimates of the intensive margin elasticity (a different concept for which quasi-experimental evidence is less abundant), two recent studies using the natural experiments of tax holidays in Iceland and Switzerland (Stefansson, 2020; Martinez *et al.*, 2021) also find intensive margin elasticities below 0.1, and that is before correction for any potential publication bias.

Publication bias does not equal cheating but arises naturally in the empirical literature even if all researchers are honest.[1] In some fields it can be addressed by the preregistration of research projects (Olken, 2015), though it is unclear whether the preregistration solution is effective outside controlled experimental research. With observational data, many researchers will write their preregistration protocols after inspecting the data or even after running preliminary analyses. Publication bias is thus a fact of life in empirical research, and it is the task of those who analyze the literature to correct for the bias. In the context of the Frisch elasticity two thresholds can potentially affect the publication probability of an estimate. First, the threshold at zero: negative estimates are economically nonsensical. Since the true elasticity cannot be negative, researchers may consider negative estimates as indicators of problems in their data or models. But negative estimates are statistically plausible given sufficient noise because few estimators of the elasticity are explicitly bounded at zero. When negative estimates are underreported, an upward bias arises in the literature since there is no psychological upper bound that would mirror and compensate for the lower bound at zero.

[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 (2001), Stanley (2005), and Stanley (2008).

Second, the threshold at the t-statistic of 1.96: two stars accompanying the regression estimate indicate that the elasticity is really far away from zero and safely in the territory prescribed by the theory. For better or worse, statistical significance has sometimes been used as an indicator of the importance of the result—and, for example, the result's usefulness for calibration. McCloskey & Ziliak (2019) provide an analogy to the Lombard effect in psychoacoustics: speakers involuntarily increase their effort with increasing noise. Similarly researchers may increase their efforts (searching through different subsets of data, models, and control variables) in response to noise in the data in order to find larger estimates and offset standard errors. With little noise and small standard errors, little or no specification search is needed to produce statistical significance. With strong noise, strong selection is required. Once again, an upward bias in the mean reported elasticity emerges as a consequence.

Our principal identification assumption in this paper is that publication bias gives rise to a positive correlation between estimates and standard errors, a correlation that does not exist in the absence of the bias. For a selection rule associated with the statistical significance threshold, the correlation arises directly from the Lombard effect. For a selection rule associated with the threshold at zero, the correlation stems from heteroskedasticity: because the true elasticity is positive, with little enough noise (and thus high enough precision) the estimates are always positive. As noise and standard errors increase, negative estimates appear from time to time but are hidden in the file drawer. Large positive estimates, which are also far away from the true value, are reported. A regression of estimates on standard errors thus yields a positive slope. (For simplicity, here we abstract from heterogeneity in the underlying elasticity for different context and individuals, which can of course affect the correlation and will be discussed and addressed later.)

The lack of correlation between estimates and standard errors in the absence of bias is a property of the methods used by the authors of the primary studies themselves. Consider, for example, the common fact that estimates are accompanied by t-statistics. Standard inference on the t-statistic makes sense only if t-statistics are symmetrically distributed. Since the t-statistic is a ratio of the point estimate to the corresponding standard error and since the symmetry property implies that the numerator and denominator are statistically independent quantities, it follows that estimates and standard errors should not be correlated. The iden-

tification assumption can be violated in economics (for example, unobserved methods choices in primary studies may systematically affect both estimates and their standard errors), and we thus relax the assumption via instrumenting the standard error by a function of the number of observations and via using a new p-uniform* technique recently developed in psychology (van Aert & van Assen, 2021) that works with the distribution of p-values instead of estimates and standard errors. The inverse of the square root of the number of observations is a natural instrument for the standard error because both quantities are correlated by the definition of the latter, and the number of observations is unlikely to be much correlated with most method choices in economics. The p-uniform* technique does not assume anything about the relation between estimates and standards errors but uses the statistical principle that the distribution of p-values is uniform at the true mean effect size.

A fact well known in the Frisch elasticity literature is that macro data tend to bring larger estimates than micro data (Chetty *et al.*, 2013). We generalize this stylized fact by showing that studies less likely to exploit genuine exogenous time variation in wages (unrelated to human capital accumulation and labor supply) are more likely to report large estimates of the elasticity. Thus the smallest elasticities are reported by studies using tax holidays, followed by other quasi-experimental studies using policy changes, often for occupations such as taxi drivers where exogenous variation in wages is more likely than for the general population. Studies using micro but non-quasi-experimental data tend to show larger elasticities, and the elasticities in macro studies are larger still. A frequent problem attributed to macro studies, but also micro studies that do not exploit policy changes staggered across several years, is the impossibility to disentangle voluntary and involuntary entries to and exits from employment. In a boom, more people can get employed simply because employers demand more labor, not just because workers choose to substitute work to the present from the past or the future in response to temporarily higher wages (Hall, 2009). We show that the ensuing identification bias is just as important as publication bias in the literature on the Frisch elasticity. After correcting for both biases we find that the literature is consistent with a mean elasticity close to zero.

The mean elasticity is important for the calibration of representative-agent models, but a very small elasticity on average does not imply that no workers substitute their labor intertemporally. Heterogeneity is important, as stressed by Attanasio *et al.* (2018), who even question

the usefulness of thinking about "the" labor supply elasticity as a unique structural parameter. We control for both underlying heterogeneity (for example age, gender, and marital status) and method heterogeneity (for example time span, data frequency, and use of instrumental variables). In total we collect 23 characteristics that reflect the context in which the estimate was obtained, and we assess which variables are effective in explaining the differences in reported elasticities. For many of the method variables no established theory exists that would mandate their inclusion in the model, but anecdotal evidence still suggests they can systematically influence the reported Frisch elasticities. Hence we face substantial model uncertainty, a natural response to which in the Bayesian framework is Bayesian model averaging (see Steel, 2020, for a detailed description). Given the number of variables and need to interpret individual marginal effects, we implement Bayesian model averaging with the dilution prior suggested by George (2010), which addresses potential collinearity. As a robustness check, we use frequentist model averaging with Mallow's weights (Hansen, 2007) and orthogonalize covariate space based on the approach of Amini & Parmeter (2012).

Our results regarding publication and identification biases are robust to controlling for heterogeneity in the estimated elasticities. We also corroborate the stylized fact that women and workers near retirement display more elastic responses than men and prime age workers. Elasticities estimated for specific industries tend to be larger than elasticities estimated for the entire economy, which is consistent with the fact that exogenous variation in wages can often be observed for occupations that are also likely to be more elastic in terms of intertemporal substitution (such as taxi drivers). Studies reporting larger estimates tend to get more citations, but it is unclear whether the correlation reflects higher quality or more convenience for calibration—larger elasticities make it often easier to match macroeconomic data (Chetty *et al.*, 2013). As the bottom line of our analysis, we use all the 723 elasticity estimates from primary studies and the model averaging exercise to compute fitted values of the elasticity conditional on a hypothetical ideal study in the literature (for example, using maximum time spans, fresh and large data, quasi-experimental design, instrumental variables to tackle measurement error, and surviving the peer review of a top five journal in economics). The mean resulting elasticity is 0.02 with the upper bound of the 95% confidence interval at 0.25. The elasticities corresponding to women and workers near retirement are around 0.1.

Three previous studies are closely related to our paper. First, Chetty *et al.* (2013) provide a detailed meta-analysis of labor supply elasticities at the extensive margin. The main part of their dataset includes Hicks elasticities; they use 6 estimates of Frisch elasticities from 6 quasi-experimental studies, while we use 723 estimates from 36 studies. Given the focus on 6 estimates, Chetty *et al.* (2013) do not examine publication bias. Second, Sokolova & Sorensen (2021) conduct a careful meta-analysis of the elasticity of labor supply to individual firms in order to analyze the extent of monopsony power. That is, Sokolova & Sorensen (2021) look at a different but related concept. They consider publication bias but focus on linear tests of the correlation between estimates and standard errors, while we employ new nonlinear tests (including Ioannidis *et al.*, 2017; Bom & Rachinger, 2019; Furukawa, 2021) and relax the exogeneity assumption using p-uniform* and via instrumenting the standard error. Third, Martinez *et al.* (2021) use the natural experiment of tax holidays in Switzerland. Given their large, high-quality dataset and the fact that the tax holidays were staggered across cantons, they are able to explore arguably exogenous time variation in net wages among the general population. It is reassuring to observe that the results of this large quasi-experiment match closely the results of a large meta-analysis: for an average person, little evidence exists of significant intertemporal substitution in labor supply at the extensive margin.

2 Estimating the Elasticity

In this section we provide a brief introduction to the Frisch elasticity and its estimation. For details on the theoretical background and empirical approaches, see Chang & Kim (2006), Keane (2011), and Attanasio *et al.* (2018); more information is also available in Section 5. Put simply, the Frisch elasticity measures how much more people want to work when their net wage increases temporarily. So the Frisch elasticity corresponds to the elasticity of substitution of labor supply. The overall effect can be disentangled into two margins: extensive (a decision whether to work at all) and intensive (a decision on how many hours to work given that one is already employed). The modern quasi-experimental literature has focused primarily on the extensive margin, and this is also the focus of our meta-analysis. In practice, the extensive margin elasticity is often computed simply as the change in the logarithm of employment rates divided by the change in the logarithm of net wages, and the latter is often instrumented. For

more context, let us start with the definition of the overall Frisch elasticity:

$$\eta = \frac{\partial h_t}{\partial w_t}\frac{w_t}{h_t}||_{\lambda}, \tag{1}$$

where h and w denote hours of work and wage, respectively. The elasticity measures the marginal change in hours worked due to the marginal change in wages while the marginal utility of lifetime wealth (λ) is held constant. Following MaCurdy (1981), in a dynamic setting without uncertainty where a temporally separable utility function (with the discount factor β), represents the household's preferences over a life cycle, the equation for estimating the elasticity can be written as:

$$\ln h_t = \alpha_i + \rho + \theta x_t + \eta \ln w_t + \varepsilon_t, \tag{2}$$

where $\alpha_i = \eta \ln \lambda$, $\rho = -\eta \ln(\beta R)$, R is the interest rate, x is a vector of characteristics affecting the household's taste for work, and ε_t is an error term.

The estimated elasticity based on this equation is usually interpreted as the aggregate response of labor supply, including both extensive and intensive margins. Assuming labor indivisibility, we can abstract from the intensive margin to address only the participation decision that operates at the extensive margin. Then the dependent variable takes a binary value, and the elasticity can be estimated by using a probit model for the participation decision. The optimal participation (employment) decision can be written as

$$h_t = \begin{cases} \bar{h}, & \text{if } w_t \geq w_t^R \\ 0, & \text{if } w_t \leq w_t^R. \end{cases} \tag{3}$$

The worker participates in the labor market and works $\bar{h}$ hours if the offered wage w_t is equal or larger than the reservation wage, w_t^R. Hence, the distribution of reservation wages plays a crucial role in determining the aggregate elasticity's magnitude at the extensive margin.

Alternatively, one can disentangle the aggregate elasticity into the intensive and extensive margins using macro data. As in Fiorito & Zanella (2012), the variance of the log of aggregate labor can be decomposed as:

$$\text{var}(\ln H_t) = \text{var}(\ln n_t) + \text{var}(\ln \bar{h}_t) + 2\,\text{cov}(\ln n_t, \ln \bar{h}_t), \tag{4}$$

where n_t is the number of employed individuals, $\bar{h}_t$ is the average number of hours worked, and aggregate labor is $H_t = n_t \bar{h}_t$. Using (4), the decomposition of aggregate Frisch elasticity can be written as

$$\eta = \frac{\text{cov}(\Delta \ln H, \Delta \ln W)}{\text{var}(\Delta \ln W)} = \frac{\text{cov}(\Delta \ln \bar{h}, \Delta \ln W)}{\text{var}(\Delta \ln W)} + \frac{\text{cov}(\Delta \ln n, \Delta \ln W)}{\text{var}(\Delta \ln W)}, \tag{5}$$

where Δ is the first-difference operator and W denotes the aggregate wage rate. The first term on the right-hand side is the intensive margin, and the second term corresponds to the extensive margin. In the extreme case where there is no heterogeneity among workers and employment is constant over the population, the extensive margin is eliminated as $\text{cov}(\Delta \ln n, \Delta \ln W) = 0$.

Apart from conventional estimation methods, some studies use nonparametric or simulation-based methods to estimate the Frisch elasticity (Erosa *et al.* 2016; Kneip *et al.* 2019). When these estimates directly capture the response of labor supply at the extensive margin, we include them as well together with controls that capture the context in which the estimates were obtained. We discuss these aspects in detail in Section 5.

3 Data

To search for empirical estimates of the elasticity we use Google Scholar because it provides a powerful full-text search. Our search procedure is described in Appendix A and conforms to the current protocol for meta-analysis (Havranek *et al.*, 2020). If the elasticity is not explicitly reported but can be calculated from the results presented in the study, we derive the elasticity and include it in our database. (In that case the standard error of the resulting elasticity is computed using the delta method.) To increase the size of the dataset available for our analysis we also include estimates from working papers. This does not help alleviate publication bias at all since working papers are intended for eventual publication and any mechanisms that lead to preference for positive or significant estimates in journal articles are also apply to working papers, as shown, for example, by Rusnak *et al.* (2013). We terminate the search on January 11, 2021, and do not add any studies beyond that date. The final sample includes 723 estimates from 36 studies (Table 1) covering a quarter century of research on labor supply elasticities. The data are available in an online appendix at `meta-analysis.cz/frisch`.

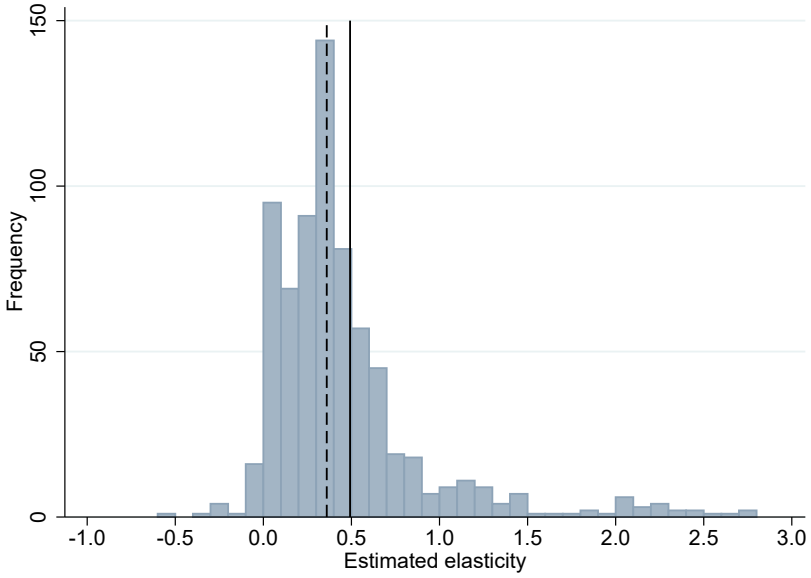
Table 1: Studies included in the meta-analysis

Attanasio *et al.* (2018)	Inoue (2015)
Bianchi *et al.* (2001)	Karabarbounis (2016)
Blundell *et al.* (2016a)	Keane & Wasi (2016)
Blundell *et al.* (2016b)	Kimmel & Kniesner (1998)
Brown (2013)	Kneip *et al.* (2019)
Caldwell (2019)	Kuroda & Yamamoto (2008)
Card & Hyslop (2005)	Looney & Singhal (2006)
Carrington (1996)	Manoli & Weber (2011)
Chang & Kim (2006)	Manoli & Weber (2016)
Chang *et al.* (2019)	Martinez *et al.* (2021)
Erosa *et al.* (2016)	Mustre del Río (2011)
Espino *et al.* (2017)	Mustre del Río (2015)
Fiorito & Zanella (2012)	Oettinger (1999)
French & Stafford (2017)	Ong (2020)
Giné *et al.* (2017)	Park (2020)
Gourio & Noual (2009)	Peterman (2016)
Gruber & Wise (1999)	Sigurdsson (2020)
Haan & Uhlendorff (2013)	Stafford (2015)

Figure 2 shows the distribution of Frisch elasticities at the extensive margin reported in the literature. The mean (0.49) is substantially larger than the median (0.36), but overall the literature appears to be quite consistent with the CBO's calibration at 0.4. We also observe that the economically impossible negative estimates sometimes appear in the literature but are very rare: a large break in the distribution of elasticities occurs at 0. That, and the skewness of the distribution with a relative abundance of elasticities above 1, is indicative of potential publication bias but little about its size and importance can be said based on a simple histogram. The dataset includes a couple of outliers on both sides of the distribution, so we winsorize the data at the 5% level. Using the outliers at their face value or omitting them from the analysis does not change our main results qualitatively.

In addition to the reported estimates and their standard errors, we collect extensive information on the context in which the estimates were obtained (22 variables in total). We control for demographic characteristics by including dummy variables reflecting whether the reported elasticity corresponds to a specific gender or age group, as well as marital status and income level. Regarding data characteristics, we control for whether the frequency of the data used is annual, quarterly, or monthly. We include controls for US data, macro data, and industry-specific data. We also include dummy variables reflecting econometric techniques (e.g., probit,

Figure 2: Estimates cluster around 0.4

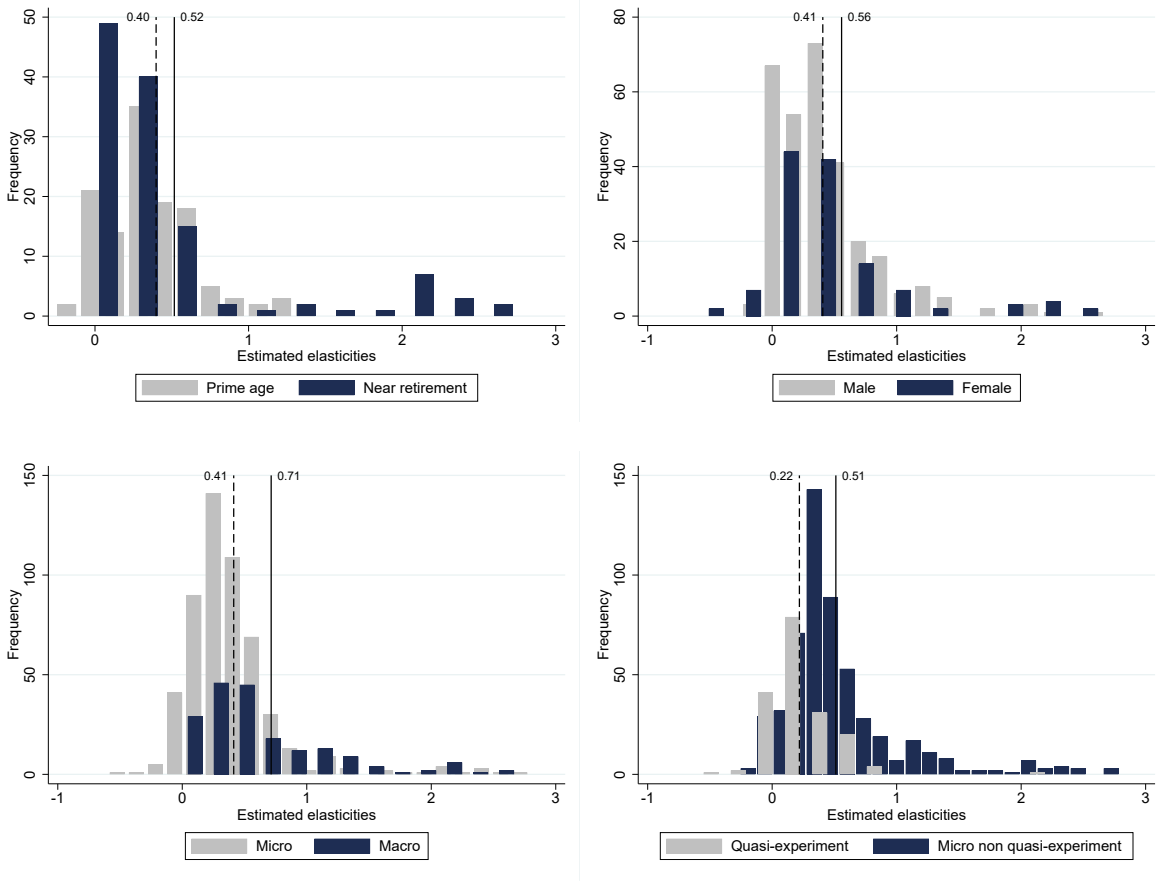


Notes: The solid line denotes the sample mean (0.49); the dashed line denotes the sample median (0.36). Estimates smaller than -1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

instrumental variables, and nonparametric methods) used in the primary studies. We control for the assumption of labor indivisibility and for quasi-experimental design. Additionally, we consider publication characteristics by controlling for study age, the number of citations, and high-quality peer-review by a top five journal in economics. Finally, we control for whether the study focuses on the Frisch elasticity or whether it reports the elasticity as a byproduct of other computations. More details on these variables are available in Section 5.

An important variable for meta-analysis is the standard error of the reported estimate. Nevertheless, for 185 estimates in our sample standard errors are not reported. To approximate standard errors, we apply the bootstrap resampling technique. We then combine the reported standard errors with those obtained from resampling. Our main results hold if we simply discard the estimates for which standard errors are not explicitly reported. Figure 3 shows four stylized facts in the data. Women and workers near retirement display larger elasticities than men and prime-age workers, which is intuitive and consistent with much of the previous literature. But the differences between women and men and between prime-age and near-retirement workers are surprisingly small, around 0.15 for gender and 0.12 for age. A larger difference arises between estimates using micro (0.41 on average) and macro data (0.71). Note that we consider

Figure 3: Stylized facts in the data



Notes: The dashed line denotes the mean elasticity for the subset mentioned first in the legend (depicted in light gray); the solid line denotes the mean for the second subset (dark). Estimates smaller than −1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

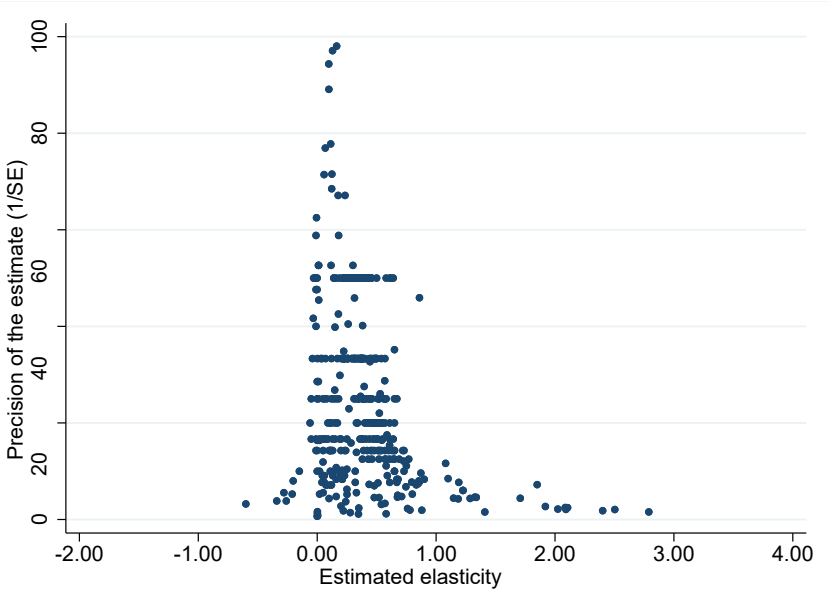
only macro estimates that explicitly try to estimate the elasticity at the extensive margin; in general, macro estimates of the aggregate Frisch elasticity tend to be even larger, and the large difference in results is well documented (Chetty *et al.*, 2013). Finally, there are is a substantial difference between micro estimates based on quasi-experimental data (0.22 on average) and non-quasi-experimental data, which use variation in taxes or wages in the absence of significant policy shifts (0.51). These stylized facts suggest both genuine differences in the elasticity (which are however modest) and the importance of proper identification. Studies more likely to exploit truly exogenous time variation in wages are also likely to report small estimates of the elasticity. But so far we have ignored the potential upward bias stemming from the selective reporting of positive and statistically significant estimates, an issue to which we turn next.

4 Publication Bias

Publication bias forces a wedge between the distribution of results obtained by researchers and the distribution of results reported by those researchers in their papers. The reported coefficients are typically larger in magnitude. To see this, consider that many effects in economics are constrained by theory to be either positive or negative. The Frisch elasticity, of course, cannot be negative, and thus negative estimates are suspicious and rarely reported. But if the true elasticity is positive and small, negative estimates will appear naturally from time to time using a method such as OLS that does not constrain the results to be positive. So a negative estimate does not necessarily imply that something is wrong with the model or the data; rather, it suggests that the underlying effect is small, estimation is imprecise, or both at the same time. In practice, the preference against negative estimates is taken a step further and leads to a preference for statistically significant positive estimates. Such estimates are sufficiently far away from the zero threshold, and statistical significance is often misused as a proxy for importance and precision. If statistical significance is the implicit or explicit goal of a researcher, it can usually be achieved by trying a sufficient number of different estimations with different methods, different subsets of data, and different control variables. At some point the researcher typically finds an estimate that is large enough to compensate the standard error and produce a t-statistic above 1.96. In both cases of selection (based on sign and on significance) an upward bias arises.

Publication bias can be assessed visually using the so-called funnel plot (Figure 4). It is a scatter plot depicting the size of the estimates on the horizontal axis and their precision on the vertical axis. Intuitively, if there is no publication bias and all studies estimate the very same parameter, the most precise estimates should be close to the underlying value of the parameter. (Sometimes the mean of the 10% most precise estimates is used as a rough estimate of the underlying effect, and Stanley *et al.*, 2010, show this simple estimator works surprisingly well. In our case the estimate derived this way is 0.26.) As precision decreases, the dispersion of estimates increases, so the figure should show an inverted funnel. An important feature of the funnel in the absence of bias is symmetry around the most precise estimates: all imprecise estimates should have the same chance of being reported. If, however, negative or small positive (and thus insignificant) imprecise estimates are underreported, the funnel becomes asymmetrical. That is what we observe in Figure 4. The most precise estimates are close to

Figure 4: The funnel plot suggests publication bias



Notes: In the absence of publication bias the plot should form a symmetrical inverted funnel. Extreme values are excluded from the figure for ease of exposition but included in all tests.

zero, but zero is also close to the bottom end of the distribution of the reported estimates. The funnel plot is a simple device developed in medical research (Egger *et al.*, 1997), where it is sometimes safe to assume homogeneity among studies, consider a linear relationship between bias and the standard error, and take reported precision at face value. But in economics all three issues are problematic, and we address them in this and the following section.

The asymmetry of the funnel plot can be tested explicitly by regressing estimates on their standard errors:

$$\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}, \tag{6}$$

where $\hat{\eta}_{ij}$ denotes the i-th estimate of the Frisch elasticity in the j-th study, $SE(\hat{\eta}_{ij})$ denotes the corresponding standard error, δ represents the size of publication bias, and η_0 can be interpreted as the peak of the funnel and thus the mean elasticity corrected for the bias (assuming that publication bias increases linearly with the standard error), an observation first made by Stanley (2005). The equation features heteroskedasticity by definition, because the explanatory variable measures the variance of the response variable. So in some applications both sides of the equations are divided by the standard error to yield a weighted least squares estimator for more efficiency. As far as we know, both the weighted and unweighted specifications were first used by

Card & Krueger (1995) and formalized by Stanley (2008) and Stanley & Doucouliagos (2012). Because most of the techniques used in the literature imply that the ratio of estimates to their standard errors has a symmetrical distribution (often a t-distribution), it follows that in the absence of publication bias there should be no correlation between the two quantities.

Table 2: Linear and nonlinear tests document publication bias

Panel A: Linear tests					
	OLS	FE	Precision	Study	IV
Standard error	1.595***	0.788***	2.222***	2.106***	2.706
(*publication bias*)	(0.262)	(0.283)	(0.484)	(0.258)	(2.103)
	[0.93, 2.20]	-	[1.23, 3.31]	[1.53, 2.56]	-
Constant	0.301***	0.370***	0.247***	0.252***	0.130*
(*mean beyond bias*)	(0.044)	(0.026)	(0.065)	(0.109)	(0.066)
	[0.12, 0.42]	-	[0.11, 0.31]	[0.13, 0.38]	-
Observations	723	723	723	723	522
Studies	36	36	36	36	23
Panel B: Nonlinear tests					
	Ioannidis *et al.* (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2021)
Effect beyond bias	0.260***	0.356***	0.207**	0.187*	0.374***
	(0.041)	(0.061)	(0.094)	(0.112)	(0.020)
Observations	723	723	723	723	723
Studies	36	36	36	36	36

Notes: Panel A presents the results of regression $\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}$, where $\hat{\eta}_{ij}$ and $SE(\hat{\eta}_{ij})$ are the i-th estimated Frisch extensive margin elasticity and its standard error reported in the j-th study. OLS = ordinary least squares. FE = study fixed effects. Precision = estimates are weighted by the inverse of their variance. Study = estimates are weighted by the inverse of the number of estimates reported per study. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error (the number of observations is not available for all studies). We cluster standard errors at the study level; if applicable, we also report 95% confidence intervals from wild bootstrap clustering in square brackets. Panel B presents the mean elasticity corrected for publication bias using nonlinear techniques. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panel A of Table 2 presents the results of estimating (6). Because most studies report more than one estimate of the elasticity, we cluster standard errors at the study level. Moreover, because the number of clusters is relatively limited (36 studies) we additionally report confidence intervals based on wild bootstrap where applicable. In addition to OLS we use study fixed effects to account for heterogeneity across studies and two weighted least squares specifications: one divides the equation by the standard error to increase efficiency, the other weights the equation by the inverse of the number of estimates reported per study in order to assign each study the same weight.

Finally, the last column of panel A addresses potential endogeneity of the standard error. The endogeneity can have at least three sources. First, the standard error is itself estimated, and this measurement error yields attenuation bias (a problem already mentioned by Stanley 2005). Second, publication selection can work on the standard error instead of the point estimate; for example, authors may choose a method that delivers statistical significance via a higher reported precision (for example, when clustering is ignored), which leads to reverse causality. Third, some method choices can influence both estimates and standard errors systematically. For example, aside from correcting a potential endogeneity problem in the point estimate, the use of instrumental variables (IV) in primary studies typically increases standard errors. While we do not see a bulletproof remedy of the endogeneity problem in meta-analysis, an appealing solution is to use the inverse of the square root of the study's number of observations as an instrument for the standard error. This is a strong instrument by the definition of the standard error (and the robust F-statistic in the first-stage regression is 11). It addresses the attenuation bias problem because the number of observations is not estimated. It addresses the reverse causality problem because a researcher cannot easily increase the number of observations just to increase significance. While some method choices can be related to the number of observations, many are independent (such as IV vs. OLS), and the instrument thus addresses the third endogeneity problem as well.

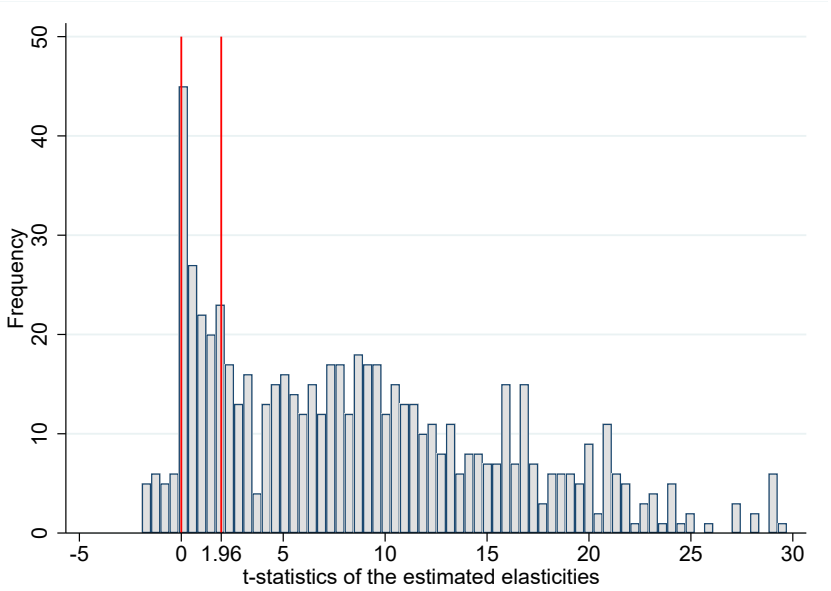
All the results in panel A of Table 2 suggest that estimates and standard errors are correlated. As expected, the IV estimate yields less precision, and even if it does not show statistical significance at conventional thresholds, the p-value is below 0.2. The point estimates of the slope coefficient range from 0.8 (fixed effects) to 2.7 (IV). Confidence intervals based on wild bootstrap range from 1 to 3. Overall, it seems that 2 is a good estimate for the slope coefficient, which translates to a strong bias. To see this, consider a case in which the true elasticity was zero. Then a slope coefficient of 2 would lead to an average reported t-statistic of 2, redrawing the inference taken from the literature. Next, as we have noted, the constant in the regression can be interpreted as the mean elasticity corrected for publication bias. The estimates range from 0.13 (IV) to 0.37 (fixed effects) with a central estimate of 0.25 and confidence intervals from 0.1 to 0.4. These results imply that publication bias exaggerates the mean elasticity about twofold, from the corrected mean of 0.25 to the reported mean of 0.49.

A problem of the funnel asymmetry test we have not yet addressed is the assumption that publication bias is a linear function of the standard error. The assumption is tenuous for small standard errors if the underlying elasticity is not zero. Consider, for example, the case when the true Frisch elasticity at the extensive margin is 0.25. When there is little noise in the data and the estimation method is sufficiently precise, the standard error will be very small: say 0.01. Then researchers will always obtain a positive and statistically significant estimate of the Frisch elasticity, and there is no reason why publication bias should arise. If the standard error is, for example, 0.02 or 0.05, the situation will not change. Publication bias will probably appear with standard errors around 0.12 and after that it may well be linearly increasing in the standard error via the mechanism described in the previous paragraphs.

Several authors have recently addressed the nonlinearity of the funnel asymmetry test, and we use a battery of these modern techniques in panel B of Table 2. First, we employ the method introduced by Ioannidis *et al.* (2017), which only uses estimates that display statistical power of at least 80% and computes the average of these estimates weighted by inverse variance. Stanley *et al.* (2017) show using Monte Carlo simulations that their technique often performs better than classical meta-analysis estimators. Second, Andrews & Kasy (2019) introduce a selection model which estimates the likelihood that negative and insignificant elasticities will be reported and then re-weights the reported estimates using the computed probabilities. Third, Bom & Rachinger (2019) assume that the relation between estimates and standard errors is nonexistent for very small standard errors and then attains a linear form discussed in the previous paragraph; the kink is estimated endogenously in the model.

Fourth, Furukawa (2021) exploits the trade-off between publication bias and variance: the most precise studies suffer less from selective reporting, but ignoring less precise studies is inefficient. His nonparametric technique estimates the share of the most precise studies that should be used for computing the corrected mean. Fifth, van Aert & van Assen (2021) do not assume anything about the correlation between estimates and standard errors, neither do they consider more precise studies to be less biased. Their technique, p-uniform*, uses the statistical principle that the distribution of p-values should be uniform at the true mean effect size. The technique is robust to heterogeneity and, by definition, also to the endogeneity of the standard error in the funnel asymmetry test.

Figure 5: Publication bias is driven by selection for positive sign, not significance



Notes: The vertical lines shown the values of t-statistics associated with changing the sign and achieving statistical significance at the 5% level.

The results of the nonlinear techniques are similar to the results reported previously for the funnel asymmetry tests. In all cases the mean corrected for publication bias is smaller than the reported mean of 0.49: estimates range from 0.19 (Furukawa, 2021) to 0.37 (van Aert & van Assen, 2021). The median estimate for the nonlinear techniques is 0.26, close to the 0.25 value in the previous panel, which suggest that the linear approximation of publication bias works well in this case. We conclude that publication bias in the literature on the Frisch elasticity at the extensive margin is substantial and likely to exaggerate the mean reported elasticity approximately twofold. As an aside, we show in Figure 5 that the bias is caused by the preference for positive sign, not statistical significance. The density of t-statistics jumps remarkably at zero, but no such jump can be seen around $t = 2$. The pattern is so clear that statistical tests are unnecessary—although caliper tests according to Gerber *et al.* (2008) and Elliott *et al.* (2021), not reported here, confirm the observation.

5 Heterogeneity

We have shown that in the literature on the Frisch elasticity publication bias is important. But what appears like publication bias can in fact be an artifact of heterogeneity. We have already

addressed heterogeneity implicitly using three estimators: the p-uniform* technique that is robust to heterogeneity, study-level fixed effects that take into account study-level differences, and an instrumental variable model that accounts for the potential endogeneity of the standard error given by, among other things, heterogeneity. In this section we model heterogeneity explicitly, and the section has three goals: first, to ascertain whether the publication bias result is robust to controlling for various aspects of estimation context; second, to identify the factors of study design that systematically influence the reported estimates; and, third, to obtain the mean elasticities conditional on various demographic characteristics and corrected for publication, identification, and other potential biases in the literature. We introduce 22 explanatory variables (in addition to the standard error) divided into four groups: characteristics of demographics, data, specification, and publication. The variables are described in Table 3.

5.1 Variables

Demographic characteristics A potentially important source of heterogeneity stems from the demographic characteristics of the samples used in primary studies. We define seven dummy variables to control for the differences in demographics. Two variables capture workers' age: although different studies use various age groups in their estimations, two groups of workers are widely highlighted in the literature. First, prime age workers between 25 and 55 years old; second, workers near retirement age (i.e., older than 55 years). Macro and micro studies disagree regarding the magnitude of the Frisch elasticity for prime age workers. Micro studies often show near-zero elasticity, while macro studies show elasticities similar to those for the whole population (Chetty *et al.*, 2013). On the other hand, workers near retirement typically exhibit a larger Frisch extensive elasticity than other age groups (e.g., Erosa *et al.*, 2016; Manoli & Weber, 2016). More than one-third of collected estimates (34%) are based on either of these groups. Elasticities based on other age groups are not commonly assessed in the literature.

Next, we codify two dummy variables denoting gender. Datasets that consist of only female workers are used for 18% of estimates, 42% of the estimates correspond to male workers only. There is a consensus in the literature that employment fluctuations in response to wages are higher among female workers than among their male counterparts. We further include a dummy variable to control for the estimates that are obtained for groups with similar income levels.

Table 3: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
Frisch elasticity	The estimated extensive margin Frisch elasticity.	0.49	0.64
Standard error	The standard error of the estimate.	0.10	0.17
Demographic characteristics			
Prime age	= 1 if the sample only consists of people between 25 and 55 years of age.	0.17	0.38
Near retirement	= 1 if the sample only consists of people older than 55.	0.17	0.38
Females only	= 1 if the sample consists of females only.	0.18	0.38
Males only	= 1 if the sample consists of males only.	0.42	0.42
Married	= 1 if the sample consists of married people only.	0.04	0.20
Single	= 1 if the sample consists of single people only.	0.03	0.18
Income	= 1 if the estimate is based on a specific income group.	0.23	0.42
Data characteristics			
Time span	The logarithm of the data time span used to estimate the elasticity.	2.22	0.89
Monthly	= 1 if the data frequency is monthly (reference category: annual).	0.02	0.15
Quarterly	= 1 if the data frequency is quarterly (reference category: annual).	0.24	0.43
Industry	= 1 if the sample consists of workers in a specific industry.	0.12	0.32
Macro	= 1 if the estimate uses aggregated data (reference category: micro).	0.26	0.44
USA	= 1 if the estimate uses data for the US.	0.67	0.47
Specification characteristics			
Indivisible labor	= 1 if the labor supply is assumed to be indivisible in the estimation framework.	0.35	0.48
Quasi-experimental	= 1 if the estimation framework uses quasi-experimental identification.	0.25	0.43
Probit	= 1 if the probit model is used for the estimate (reference category: OLS).	0.05	0.22
Non-parametric	= 1 if non-parametric simulation-based methods are used (reference category: OLS).	0.37	0.48
IV	= 1 if instrumental variable methods are used for the estimate (reference category: OLS).	0.14	0.35
Publication characteristics			
Publication year	The logarithm of the publication year of the study.	3.46	0.20
Top journal	= 1 if the estimate is published in a top five journal in economics.	0.22	0.41
Citations	The logarithm of the number of per-year citations of the study, according to Google Scholar.	1.75	1.06
Byproduct	= 1 if the information reported in the study allows for the computation of the elasticity but the elasticity is not interpreted in the paper.	0.04	0.20

Notes: SD = standard deviation. The table excludes the definition and summary statistics of the reference categories, which are omitted from the regressions.

Almost 23% of estimates are based on a specific homogeneous income group. Finally, two dummy variables control for the marital status of the people examined. Only 4% of estimates correspond to married workers only, and 3% for single workers only. Although we collect two extra dummy variables that capture elasticities computed for workers without children and self-employed workers, these subsamples are used rarely in the literature and the corresponding variables have very little variance. Hence we exclude them from the analysis.

Data characteristics The second category of variables covers the characteristics of the data used in estimations. We introduce a variable reflecting the time span of the data. Moreover, two dummy variables control for data frequency. We use annual data as the reference category since more than 74% of estimates employ annual data; as noted by Martinez *et al.* (2021), annual frequency is the relevant time frame for business cycle analysis. The dummy variable "Industry" controls for the fact whether the estimate uses data from a specific industry. More than 67% of the estimates utilize datasets relevant to the US, including The Panel Study of Income Dynamics and the National Longitudinal Survey of Youth. We thus add a dummy variable for the use of US data. The majority of the estimates (74%) use individual-level data, while others use aggregate-level (macro) data. We use the former as the baseline category and define a dummy variable for the latter.

Specification characteristics We use five variables to control for the specification of primary studies. The first variable equals one if the estimate assumes the indivisibility of labor. In this case, since people can either work full-time or be unemployed, all labor fluctuations appear at the extensive margin. Slightly more than a third of the estimates employ the indivisible labor assumption. Next, quasi-experimental estimates account for one-fourth of all estimates in the primary studies. Quasi-experimental studies yield a mean estimate of 0.22, substantially smaller than the mean estimate from the remainder of the studies (0.58). Within quasi-experimental studies, some are arguably even better specified, especially those that use data on tax holidays from Iceland and Switzerland (Stefansson, 2020; Martinez *et al.*, 2021), and thus have the best chance to exploit exogenous time variation in net wages. But because there are few such studies, we cannot meaningfully create a separate dummy for them. Additionally, three additional dummy variables control for the potential effect of econometric techniques used in estimating

elasticities. The baseline category is OLS, as researchers use it to estimate more than 45% of estimates. Probit models are used only in 5% of estimates, while the instrumental variables and non-parametric methods are used in 14% and 37% of estimates, respectively.

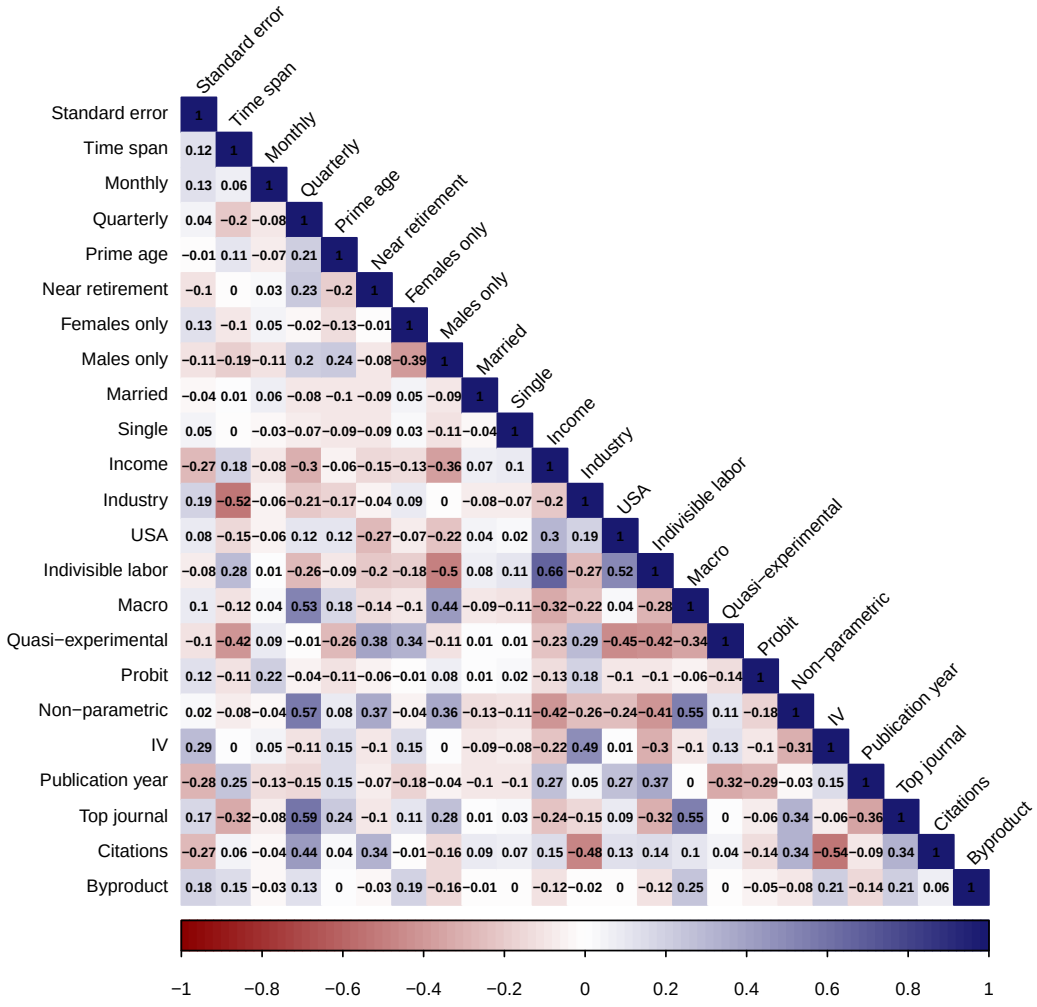
Publication characteristics The last category of variables attempts to capture quality not reflected by the variables introduced above. First, we account for the publication year of the study—ceteris paribus, more recent studies are likely to bring improvements in data and methods that might be difficult to pin down explicitly. The next variable reflects the logarithm of the number of per-year citations of the study according to Google Scholar. We expect studies of higher quality to be quoted more frequently, but on the other hand the number of citations can also be correlated with the size of the elasticity simply because structural macro models need larger estimates of the elasticity for calibration. Next, to account for high-quality peer review, we include a dummy variable for the case when the study is published in one the top five journals. Finally, we create a variable that equals one if the estimate is either a byproduct of different analyses in the study. For example, Carrington (1996) and Brown (2013) do not directly report the estimated Frisch extensive elasticity, while Chang & Kim (2006) report the estimated Frisch extensive elasticity as a supplement.

Figure 6 shows that correlations among the variables are not extensive. The largest correlation coefficient is 0.66, and all variance-inflation factors are below 10. But given the number of explanatory variables and need to interpret individual marginal effects in regressions, we use a method that takes potential collinearity into account (the dilution prior). Figure 6 shows some stylized facts of the literature: for example, quasi-experimental studies tend to have relatively short time spans and are often conducted using non-US data, standard errors tend to be larger when instrumental variables are employed and smaller when a homogeneous datasets (based on individuals' income) is used, macro studies often use data at the quarterly frequency, and time spans used in studies have been increasing recently.

5.2 Estimation

The intuitive approach to model heterogeneity is to regress the reported elasticities on all the variables introduced above. But that is incorrect because it ignores model uncertainty: while we

Figure 6: Correlations among explanatory variables are modest



Notes: The figure shows Pearson correlation coefficients for the explanatory variables described in Table 3.

want to control for all of the variables introduced above, we are not sure that all of them belong to the underlying model. A simple OLS regression would result in inefficient estimates. In fact, a regression with all the variables included is only one of many millions of potential models. A natural solution to model uncertainty in the Bayesian setting is Bayesian model averaging (BMA). Using all the possible subsets of explanatory variables (i.e., 2^k, where k is the number of explanatory variables), BMA runs numerous regression models. Analogous to the information criteria in frequentist econometrics, posterior model probability (PMP) is assigned to each model. PMP assesses the performance of a model (in terms of fit and parsimony) compared to other models. BMA uses weights based on PMPs to construct a weighted average over the estimated coefficients across all the models. Furthermore, posterior inclusion probability (PIP)

is constructed for each variable and indicates the sum of posterior model probabilities of the models in which the variable is included. Further details on BMA can be found in, e.g., Raftery *et al.* (1997) and Eicher *et al.* (2011).

Estimating 2^{23} models would take days using a standard personal computer. Hence, we apply the Markov chain Monte Carlo algorithm (Madigan & York, 1995), which goes through the models with the highest posterior model probabilities. We implement BMA using the `bms` package developed by Zeugner & Feldkircher (2015). In the baseline specification we employ the dilution prior suggested by George (2010), which takes into account the collinearity of the variables included in each model. The prior multiplies the model probabilities by the determinant of the correlation matrix of the variables. Higher collinearity means that the determinant is closer to zero, which results in a model with little weight. Following Eicher *et al.* (2011), we also use the unit information prior (UIP) for Zellner's g-prior, in which the prior that all regression parameters are zero has the same weight as one observation in the data. In addition, we run a frequentist check, which is a hybrid frequentist-Bayesian model that only includes variables with PIPs higher than 0.75 obtained from the baseline BMA specification. We then estimate the model using OLS and cluster standard errors at the study level.

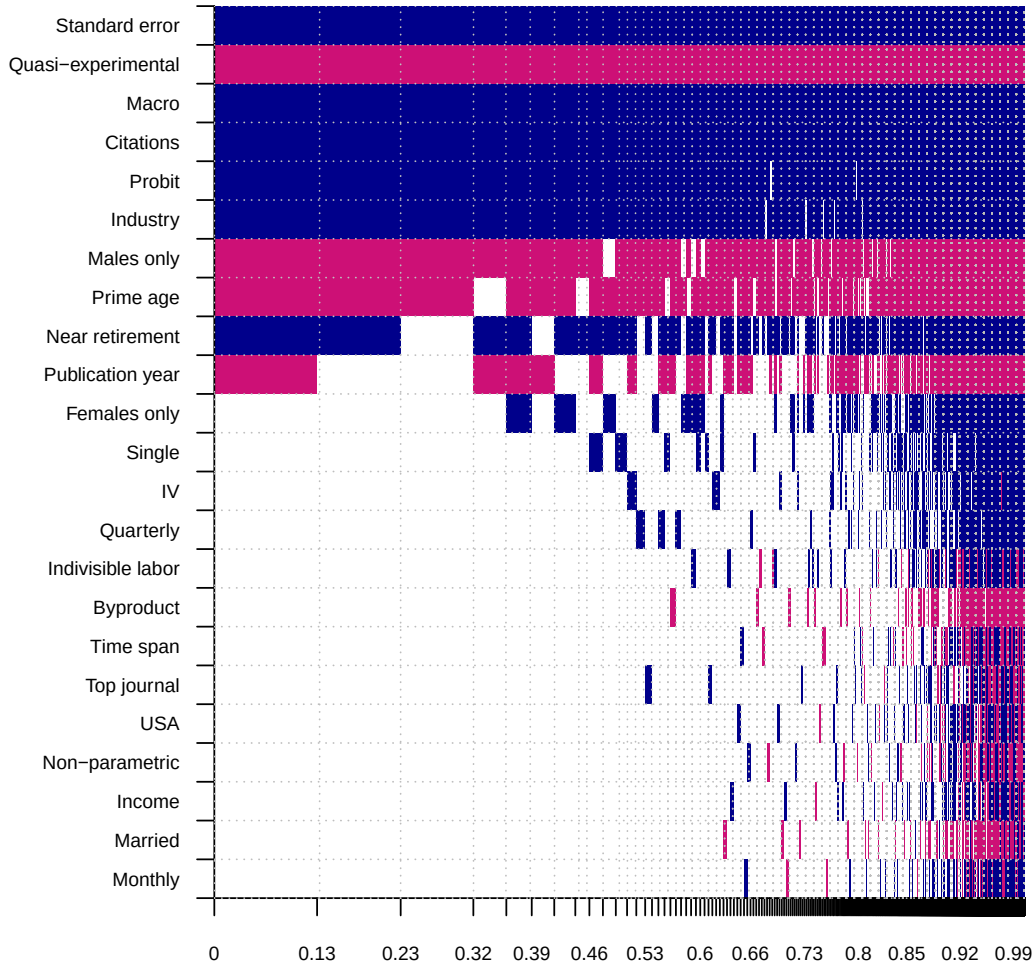
5.3 Results

Figure 7 illustrates the results of Bayesian model averaging. Each column represents an individual regression model, and the models are sorted on the horizontal axis by their posterior model probabilities from the best model on the left. The vertical axis shows the explanatory variables listed in the descending order of their posterior inclusion probabilities. The blue color (darker in grayscale) indicates that the corresponding coefficient is positive, while the red color (lighter in grayscale) denotes the negative sign of the coefficient. A blank cell means that the corresponding variable is not included in the model. At first glance, Figure 7 indicates that 10 variables seem to be systematically important in explaining the heterogeneity of the reported elasticities: these variables have high PIPs and robust signs across regression models.

Table 4 presents the numerical results of Bayesian model averaging. The left panel reports the posterior inclusion probability, posterior mean, and posterior standard deviation for each explanatory variable's regression coefficient. Excluding the intercept, three variables have PIP

equal to 1, indicating that they are *decisive* variables (in the classification of Raftery *et al.* 1997); three variables are *strong* as their PIPs are between 0.95 and 0.99, and two can be labeled as *substantial* with PIPs more than 0.75 but lower than 0.95. Moreover, there are two variables with PIPs between 0.50 and 0.75, indicating their *weak* effect. The right panel of Table 4 shows the results of OLS, including the variables with PIP 0.75 and higher. The estimated coefficients in both panels have the same sign and similar magnitude, and apart from one variable, they display the same statistical importance (PIP in BMA and its frequentist equivalent, p-value). So the results of the frequentist check are consistent with the baseline BMA.

Figure 7: Model inclusion in Bayesian model averaging



Notes: The response variable is the reported estimate of the Frisch elasticity of labor supply at the extensive margin. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on the unit information prior (UIP) recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010), which takes collinearity into account. Blue color (darker in grayscale) = the variable has a positive estimated sign. Red color (lighter in grayscale) = the variable has a negative estimated sign. No color = the variable is excluded from the given model. Table 3 presents a detailed description of all variables. The numerical results are reported in Table 4.

Table 4: Why do estimates of the elasticity vary?

	Bayesian Model Averaging			Frequentist Check (OLS)		
Variable	Post. Mean	Post. SD	PIP	Coeff.	S.E.	p-val.
Intercept	0.558	N.A.	1.000	0.210	0.048	0.000
Standard error	1.331	0.146	1.000	1.407	0.135	0.000
Demographic characteristics						
Prime age	-0.092	0.048	0.864	-0.103	0.035	0.005
Near retirement	0.082	0.057	0.749	0.093	0.071	0.200
Females only	0.020	0.041	0.243			
Males only	-0.102	0.043	0.911	-0.110	0.055	0.054
Married	-0.001	0.011	0.032			
Single	0.012	0.040	0.118			
Income	0.000	0.007	0.035			
Data characteristics						
Time span	0.000	0.006	0.044			
Monthly	0.001	0.015	0.031			
Quarterly	0.004	0.017	0.071			
Industry	0.174	0.058	0.968	0.160	0.069	0.027
Macro	0.197	0.036	1.000	0.207	0.050	0.000
USA	0.001	0.010	0.040			
Specification characteristics						
Indivisible labor	0.003	0.021	0.068			
Quasi-experimental	-0.293	0.049	1.000	-0.270	0.049	0.000
Probit	0.244	0.068	0.989	0.277	0.131	0.041
Non-parametric	0.000	0.009	0.038			
IV	0.005	0.021	0.075			
Publication characteristics						
Publication year	-0.099	0.114	0.497			
Top journal	0.001	0.012	0.043			
Citations	0.076	0.016	0.999	0.076	0.020	0.001
Byproduct	-0.003	0.018	0.048			
Observations	723			723		
Studies	36			36		

Notes: The response variable is the Frisch elasticity of labor supply at the extensive margin. S.D. = standard deviation, PIP = Posterior inclusion probability, S.E. = standard error. The left-hand panel applies BMA based on the UIP g-prior and the dilution prior (Eicher *et al.* 2011; George 2010). The right-hand panel reports a frequentist check using OLS, which includes variables with PIPs higher than 0.75 in BMA. Standard errors in the frequentist check are clustered at the study level. Table 3 presents a detailed description of all the variables.

The first important conclusion from Bayesian model averaging is that our result concerning publication bias remains robust even when we explicitly take into account the context in which the elasticity is estimated by adding extra 22 explanatory variables to our regression model. The effect of publication bias in BMA results is in line with the findings reported in the previous section. BMA results show that publication bias exaggerates the estimated Frisch extensive elasticities, confirming that the significant correlation between standard errors and estimates is not due to omitted aspects of demographics, data, specification, and publication.

Demographics. We find that demographic characteristics affect the estimates of the Frisch extensive elasticity in different respects. First, the estimates for men tend to be smaller than those for women. Our results also suggest that estimates of the elasticity for workers near retirement are systematically larger than elasticities for other age groups, especially prime age workers. The findings confirm the patterns in the literature shown earlier in Figure 3 and are also in line with the consensus in the literature. Card & Hyslop (2005), Keane (2011), and Keane & Rogerson (2015), for instance, document that women and workers near retirement display relatively large elasticities since they are less attached to the labor market compared to other demographic groups. Our results do not suggest that marital status can explain the variation in the reported estimates. Similarly, we find no evidence that studies using homogeneous groups of workers with similar income tend to find elasticities systematically different from the rest of the literature.

Data characteristics. Our results indicate no systematic effect of the time span and data frequency used in the primary study on the reported elasticity. We do not find evidence that the US-based estimates are systematically different from estimates reported for other countries. In contrast, elasticities obtained from macro data tend to by systematically larger than elasticities obtained from micro data, which is a stylized fact well known in the literature (Chetty *et al.*, 2013). In addition, our analysis suggests that there is a systematic relationship between industry-specific data and reported estimates of the Frisch extensive elasticity. Industry-specific estimates are systematically larger than estimates that are not associated with particular industries, perhaps because exogenous time variation in net wages is often available for groups

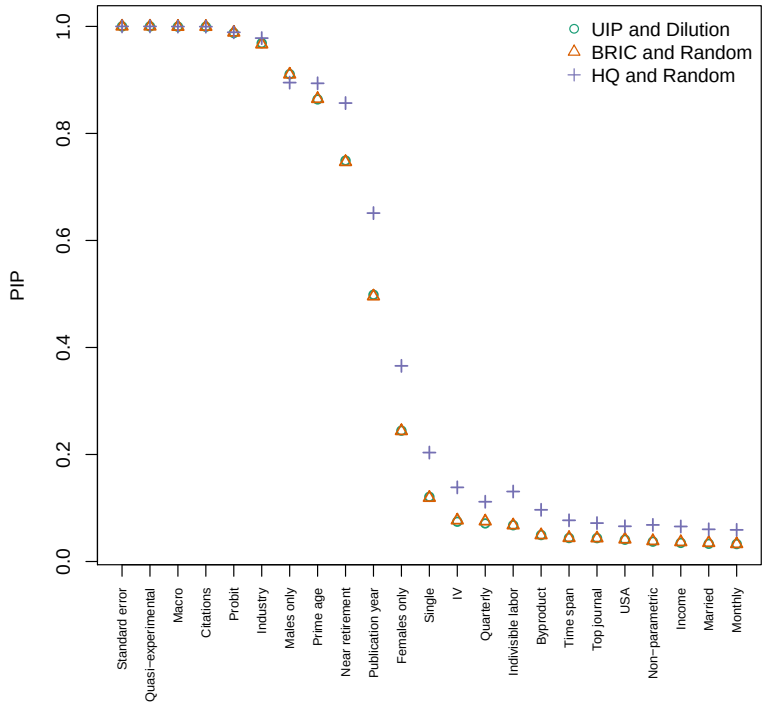
that are also likely to display more intertemporal substitution (such as fishermen, taxi drivers, and bike messengers).

Specifications. We find that assuming labor indivisibility is not systematically related to the size of the elasticity. The result contrasts a part of the macro literature, initiated by Hansen (1985) and Rogerson (1988), highlighting the importance of indivisible labor supply in determining the Frisch extensive elasticity. We find little evidence that either IV or non-parametric techniques used in estimating the elasticity affect the results systematically. On the other hand, elasticities estimated by the probit technique tend to be systematically larger. Finally and importantly, our results suggest that the quasi-experimental research design is a key factor for explaining the heterogeneity in the literature. Studies that do not follow the quasi-experimental approach tend to report larger estimates by 0.3 on average. This finding corroborates the pattern depicted earlier in Figure 3.

Publication characteristics. Regarding potentially unobserved aspects of quality, our results suggest little systematic effects of publication year, publication in a top-five journal, and focus of the study (whether the study estimates the Frisch elasticity explicitly or focuses on a different exercise and derives the elasticity only as a byproduct). In contrast, the number of citations is robustly associated with the reported elasticities, and the correlation is positive. The finding is interesting but we are unable to establish causality in this case. On the one hand, perhaps citations really serve as a good proxy for unobserved quality, and so better studies do produce larger elasticities. On the other hand, some studies can be cited more often precisely because they report larger elasticities, since larger elasticities are more convenient for the calibration of many structural macro models.

In addition to the baseline BMA we conduct a series of robustness checks. First, we employ alternative model priors and parameter g-priors. We apply the beta-binomial random model prior, which gives an equal prior probability to each model size (Ley & Steel, 2009). We also use the BRIC g-prior suggested by Fernandez *et al.* (2001) together with the HQ prior. Figure 8 depicts how the posterior inclusion probabilities change when we change priors: the changes in PIPs are small. The detailed results obtained from alternative BMA settings are presented in Appendix B. As another sensitivity check, we give each study the same importance in BMA by

Figure 8: Posterior inclusion probabilities hold across different priors



Notes: UIP and Dilution = priors according to Eicher *et al.* (2011) and George (2010). BRIC and Random = the benchmark g-prior for parameters with the beta-binomial model prior (each model size has equal prior probability). The HQ prior asymptotically mimics the Hannan-Quinn criterion.

using as a weight the inverse of the number of estimates reported per study. Next, we run BMA without the standard error variable, which is correlated with some of the method choices. The results of these two robustness checks are available in Table B3. Finally, we apply frequentist model averaging (FMA), which does not need priors. We use Mallow's weights (Hansen, 2007) and the orthogonalization of covariate space suggested by Amini & Parmeter (2012). Table B4 reports the results of FMA exercise. The robustness checks corroborate our main results, but it is worth noting that FMA finds a systematic effect of the use of instrumental variables in primary studies (although the effect is only significant at the 10% level). According to FMA, IV estimates are on average 0.1 larger than OLS estimates, which is consistent with a modest attenuation bias in the literature.

5.4 Implied Elasticities

As the bottom line of our analysis we compute the Frisch extensive margin elasticity implied by the literature and conditional on the absence of publication bias, identification bias, and

other estimation problems. In other words, we create a hypothetical study that uses all information and estimates reported in the literature but puts more weight on the aspects of data and methodology that are arguably preferable. Such a "best-practice" exercise is inevitably subjective, because different researchers have different opinions on what constitutes best practice. So we try to be conservative and choose best practice values only for a couple of the most important aspects of study design, while remaining agnostic about the rest. Aside from our definition of best practice we use an alternative definition which relies on the design of an influential study, Blundell *et al.* (2016a). In practice, we use the results of model averaging and compute fitted values of the Frisch elasticity when specific values of the 23 variables are plugged in. When we have no preference about the particular aspect of study design, we plug in the sample mean; otherwise, we plug in the preferred value (for example, we plug in 1 for the dummy variable corresponding to quasi-experimental design). In order to compute confidence intervals, we use the results of frequentist model averaging, but Bayesian model averaging gives similar point estimates.

To correct for publication bias, we plug in zero for the standard error—in other words, we condition the estimation of the implied elasticity on maximum precision in primary studies. While the linear model of publication bias with an exogenous standard error is simplistic, we have shown earlier that it works well in the case of the Frisch elasticity and yields results similar to nonlinear techniques and techniques that allow for endogeneity in the standard error. We prefer longer time spans in primary studies and plug in the sample maximum for the corresponding variable. We prefer annual data and so plug in zeros for monthly and quarterly dummies; as noted by Martinez *et al.* (2021), annual frequency is the relevant time frame for business cycle analysis. For the overall estimate we also prefer samples of general population, so we plug in zeros for female, male, prime-age, and near-retirement dummies. We also prefer when the elasticity is computed for the entire economy, not an individual industry. We prefer micro, quasi-experimental data. We plug in 1 for instrumental variable estimation in order to take into account attenuation bias and other potential biases related to endogeneity, at least to the extent that the instrumental variables used in primary studies can address the biases. We prefer studies published recently and put more weight on high-quality peer-review (proxied by publication in a top five journal in economics). Finally, we prefer when the study focuses

Table 5: Mean elasticities implied by the literature are close to zero

	Subjective best practice	Blundell *et al.* (2016a)
Overall	0.02	−0.03
	(−0.22, 0.25)	(−0.25, 0.18)
Near retirement	0.14	0.09
	(−0.14, 0.42)	(−0.14, 0.31)
Women	0.12	0.07
	(−0.08, 0.32)	(−0.10, 0.24)

Notes: The table shows elasticities implied by the literature and conditional on selected characteristics of demographics, specification, data, and publication. The benchmark estimate in the first row corresponds to the overall mean elasticity; the next two rows show estimates for workers near retirement and women. In the first column we construct a definition of best practice based on our reading of the literature. For the computation we use the results of frequentist model averaging and compute fitted values conditional on the definition of best practice (for example, we use 0 for the standard error in order to correct for publication bias and 1 for the quasi-experimental dummy variable in order to put more weight on quasi-experimental results). In the second column we do not define best practice ourselves but use the characteristics used by Blundell *et al.* (2016a). The 95% confidence intervals are reported in parentheses.

directly on the elasticity and does not compute the elasticity merely as a byproduct of another exercise. All other variables are set to their sample means.

Table 5 shows the results. The first column presents our subjective best practice defined in the previous paragraph. In the second column we conduct a similar exercise but instead of selecting aspects of best practice subjectively we choose the aspects of the baseline estimation in Blundell *et al.* (2016a). Both approaches give similar results: the point estimate of the mean elasticity is about zero, and the mean is below 0.25 with 95% confidence. The next two rows show that the implied elasticities are a bit larger for women and workers near retirement, but even here the elasticities are smaller than 0.42 with 95% confidence. The CBO's calibration of the mean Frisch elasticity at 0.4 is thus inconsistent with the literature.

6 Concluding Remarks

A general implication of our results is that it is wrong to calibrate a parameter of a structural macro model, or indeed any policy, based on the mean estimate of that parameter reported in the micro literature. The reported mean is an extremely poor reflection of the underlying parameter. Heterogeneity is one problem, but to calibrate a representative-agent model one still needs a representative value. The main issue is publication bias, which in our case exaggerates the mean reported estimate twofold. Remarkably, the same degree of exaggeration due to publication bias has been found by Ioannidis *et al.* (2017) for the empirical economics literature as a whole. What

is more, the same exaggeration has also been shown by preregistered replications of estimations in economics and psychology by Open Science Collaboration (2015) and Camerer *et al.* (2018). So a plausible rule of thumb, in the absence of other useful information, is to calibrate a parameter at half the mean value reported in the literature. (Among the few researchers who specialize on meta-analysis in economics, the rule has been known as the Paldam rule after the Danish economist Martin Paldam who, half-jokingly, suggested it at a conference five years ago.) But we also show that identification problems can be, on average, just as important as publication bias. No simple rule can address identification bias, and in the absence of a detailed and careful meta-analysis it can well be better to focus on a recent, large, and well-identified primary study instead of the mean of the entire literature. We argue that for the Frisch elasticity Martinez *et al.* (2021) provide such a study, and their results are consistent with our large meta-analysis: intertemporal substitution in labor supply is weak at best.

If a high-quality primary study can serve as a good guide for calibration or policy, why bother with a laborious and time-consuming meta-analysis? Publication bias is not a problem of literature surveys exclusively—it can affect the results reported in any primary study, and without meta-analysis we have no idea about the extent of the potential problem. In contrast to individual studies and narrative surveys, meta-analysis can address both publication and identification biases at the same time. A comparison with a large, high-quality primary study provides an important robustness check. The dataset of Martinez *et al.* (2021) is so large that they can identify statistical significance even for intensive margin elasticities as small as 0.02. Given such great statistical power and hence high estimation precision, it would be extremely difficult to produce large estimates of the elasticity even if the authors were inclined to do so. But still the data on this natural experiment correspond to a small European country, and without a detailed meta-analysis it is unclear whether these results are valid externally. We now argue with confidence that they are.

An important problem we cannot fully address is potential attenuation bias, the "iron law of econometrics" (Hausman, 2001). Wages and taxes are measured with an error, especially in surveys. If the measurement error is large and the authors of primary studies do not address it adequately, our results, and the results of Martinez *et al.* (2021), understate the strength of intertemporal substitution. A crude way how to evaluate the extent of attenuation bias is

to compare estimates obtained using instrumental variables with those obtained using OLS. If the instruments are valid and the measurement error in instruments is not related to the measurement error in wages or taxes, the difference between IV and OLS estimates indicates the size of attenuation bias—though together with other potential endogeneity biases. We find little systematic differences between both types of estimates. In Bayesian model averaging the difference is zero. In frequentist model averaging IV estimates tend to be 0.1 larger than OLS estimates, other things being equal. So our analysis is consistent with a small attenuation bias that nevertheless does not affect the results qualitatively.

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 & Utrecht 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. 2766–94.

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.

Attanasio, O., P. Levell, H. Low, & V. Sánchez-Marcos (2018): "Aggregating elasticities: intensive and extensive margins of women's labor supply." *Econometrica* **86(6)**: pp. 2049–2082.

Bianchi, M., B. R. Gudmundsson, & G. Zoega (2001): "Iceland's natural experiment in supply-side economics." *American Economic Review* **91(5)**: pp. 1564–1579.

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

Blundell, R., M. Costa Dias, C. Meghir, & J. Shaw (2016a): "Female labor supply, human capital, and welfare reform." *Econometrica* **84(5)**: pp. 1705–1753.

Blundell, R., L. Pistaferri, & I. Saporta-Eksten (2016b): "Consumption inequality and family labor supply." *American Economic Review* **106(2)**: pp. 387–435.

Bom, P. R. & H. Rachinger (2019): "A kinked meta-regression model for publication bias correction." *Research Synthesis Methods* **10(4)**: pp. 497–514.

Brodeur, A., N. Cook, & A. Heyes (2020): "Methods Matter: P-Hacking and Causal Inference in Economics." *American Economic Review* **110(11)**: pp. 3634–3660.

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

Brown, K. M. (2013): "The link between pensions and retirement timing: lessons from California teachers." *Journal of Public Economics* **98**: pp. 1–14.

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

Caldwell, S. C. (2019): *Essays on imperfect competition in the labor market*. Ph.D. thesis, Massachusetts Institute of Technology.

Camerer, C. F., A. Dreber, F. Holzmeister, T. H. Ho, J. Huber, M. Johannesson, M. Kircher, G. N. G, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell,

A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E. J. Wagenmakers, & H. Wu (2018): "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* **2**: pp. 637–644.

Card, D. & D. R. Hyslop (2005): "Estimating the effects of a time-limited earnings subsidy for welfare-leavers." *Econometrica* **73(6)**: pp. 1723–1770.

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." *American Economic Review* **85(2)**: pp. 238–243.

Carrington, W. J. (1996): "The Alaskan labor market during the pipeline era." *Journal of Political Economy* **104(1)**: pp. 186–218.

Chang, Y. & S.-B. Kim (2006): "From individual to aggregate labor supply: a quantitative analysis based on a heterogeneous agent macroeconomy." *International Economic Review* **47(1)**: pp. 1–27.

Chang, Y., S.-B. Kim, K. Kwon, & R. Rogerson (2019): "2018 Klein lecture: individual and aggregate labor supply in heterogeneous agent economies with intensive and extensive Margins." *International Economic Review* **60(1)**: pp. 3–24.

Chetty, R., A. Guren, D. Manoli, & A. Weber (2013): "Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities." *NBER Macroeconomics Annual* **27(1)**: pp. 1–56.

Christensen, G. & E. Miguel (2018): "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* **56(3)**: pp. 920–980.

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

Egger, M., G. D. Smith, M. Schneider, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *British Medical Journal* **315(7109)**: pp. 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)**.

Erosa, A., L. Fuster, & G. Kambourov (2016): "Towards a micro-founded theory of aggregate labour supply." *The Review of Economic Studies* **83(3)**: pp. 1001–1039.

Espino, A., F. Isabella, M. Leites, & A. Machado (2017): "Do women have different labor supply behaviors? evidence based on educational groups in Uruguay." *Feminist Economics* **23(4)**: pp. 143–169.

Fernandez, C., E. Ley, & M. F. Steel (2001): "Benchmark priors for Bayesian model averaging." *Journal of Econometrics* **100(2)**: pp. 381–427.

Fiorito, R. & G. Zanella (2012): "The anatomy of the aggregate labor supply elasticity." *Review of Economic Dynamics* **15(2)**: pp. 171–187.

French, S. & T. Stafford (2017): "Returns to experience and the elasticity of labor supply." *Working paper 2017-15*, UNSW Business School.

Furukawa, C. (2021): "Publication bias under aggregation frictions: from communication model to new correction method." *Working paper*, MIT, mimeo.

Galí, J. (2015): *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, second edition.

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 *et al.* (2008): "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.

Giné, X., M. Martinez-Bravo, & M. Vidal-Fernández (2017): "Are labor supply decisions consistent with neoclassical preferences? evidence from Indian boat owners." *Journal of Economic Behavior & Organization* **142**: pp. 331 – 347.

Gourio, F. & P.-A. Noual (2009): "The marginal worker and the aggregate elasticity of labor supply."

Boston University Dept. of Economics Working Papers Series, WP2006-009 .

Gruber, J. & D. A. Wise (1999): *Social Security and Retirement around the World*. University of Chicago Press.

Haan, P. & A. Uhlendorff (2013): "Intertemporal labor supply and involuntary unemployment." *Empirical Economics* **44(2)**: pp. 661–683.

Hall, R. E. (2009): "Reconciling Cyclical Movements in the Marginal Value of Time and the Marginal Product of Labor." *Journal of Political Economy* **117(2)**: pp. 281–323.

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

Hansen, G. D. (1985): "Indivisible labor and the business cycle." *Journal of Monetary Economics* **16(3)**: pp. 309–327.

Hausman, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives* **15(4)**: pp. 57–67.

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.

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

Inoue, Y. (2015): "Intensive and extensive margins of Japanese male and female workers- evidence from the tax policy Reform in Japan." *Working paper*, Panel Data Research Center at Keio University.

Ioannidis, J. P., T. D. Stanley, & H. Doucouliagos (2017): "The power of bias in economics research." *The Economic Journal* **127(605)**: pp. F236–F265.

Karabarbounis, M. (2016): "A road map for efficiently taxing heterogeneous agents." *American Economic Journal: Macroeconomics* **8(2)**: pp. 182–214.

Keane, M. & R. Rogerson (2015): "Reconciling micro and macro labor supply elasticities: A structural perspective." *Annual Review of Economics* **7(1)**: pp. 89–117.

Keane, M. P. (2011): "Labor supply and taxes: A survey." *Journal of Economic Literature* **49(4)**: pp. 961–1075.

Keane, M. P. & N. Wasi (2016): "Labour supply: the roles of human capital and the extensive margin." *The Economic Journal* **126(592)**: pp. 578–617.

Kimmel, J. & T. J. Kniesner (1998): "New evidence on labor supply: employment versus hours elasticities by sex and marital status." *Journal of Monetary Economics* **42(2)**: pp. 289–301.

Kneip, A., M. Merz, & L. Storjohann (2019): "Aggregation and labor supply elasticities." *Journal of the European Economic Association* **18(5)**: pp. 2315–2358.

Kuroda, S. & I. Yamamoto (2008): "Estimating Frisch labor supply elasticity in Japan." *Journal of the Japanese and International Economies* **22(4)**: pp. 566–585.

Ley, E. & M. F. Steel (2009): "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* pp. 651–674.

Looney, A. & M. Singhal (2006): "The effect of anticipated tax changes on intertemporal labor supply and the realization of taxable income." *Working paper 12417*, National Bureau of Economic Research.

MaCurdy, T. E. (1981): "An empirical model of labor supply in a life-cycle setting." *Journal of Political Economy* **89(6)**: pp. 1059–1085.

Madigan, D. & J. York (1995): "Bayesian graphical models for discrete data." *International Statistical Review* **63(2)**: pp. 215–232.

Manoli, D. & A. Weber (2011): "Nonparametric Evidence on the Effects of Retirement Benefits on Labor Force Participation Decisions." *Working Papers, Center for Retirement Research at Boston College wp2011-24*, Center for Retirement Research.

Manoli, D. & A. Weber (2016): "Nonparametric evidence on the effects of financial incentives on retirement decisions." *American Economic Journal: Economic Policy* **8(4)**: pp. 160–82.

Martinez, I. Z., E. Saez, & M. Siegenthaler (2021): "Intertemporal Labor Supply Substitution? Evidence from the Swiss Income Tax Holidays." *Amer-*

ican Economic Review **111(2)**: pp. 506–546.

McCloskey, D. N. & 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.

Oettinger, G. S. (1999): "An empirical analysis of the daily labor supply of stadium venors." *Journal of Political Economy* **107(2)**: pp. 360–392.

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

Ong, P. (2020): "The effect of child support on labor supply: An estimate of the Frisch elasticity." *Working paper*, National University of Singapore.

Open Science Collaboration (2015): "Estimating the reproducibility of psychological science." *Science* **349(6251)**: p. aac4716.

Park, C. (2020): "Consumption, reservation wages, and aggregate labor supply." *Review of Economic Dynamics* **37(1)**: pp. 54–80.

Peterman, W. B. (2016): "Reconciling micro and macro estimates of the Frisch labor supply elasticity." *Economic Inquiry* **54(1)**: pp. 100–120.

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.

Mustre del Río, J. (2011): "The aggregate implications of individual labor supply heterogeneity." *Working paper*, Federal Research Bank of Kansas City, Research Division.

Mustre del Río, J. (2015): "Wealth and Labor Supply Heterogeneity." *Review of Economic Dynamics* **18(3)**: pp. 619–634.

Rogerson, R. (1988): "Indivisible labor, lotteries and equilibrium." *Journal of Monetary Economics* **21(1)**: pp. 3–16.

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.

Sigurdsson, J. (2020): "Labor supply responses and adjustment frictions: a tax-free year in Iceland." *Working paper*, Norwegian School of Economics.

Sokolova, A. & T. Sorensen (2021): "Monopsony in Labor Markets: A Meta-Analysis." *ILR Review* **74(1)**: pp. 27–55.

Stafford, T. M. (2015): "What do fishermen tell us that taxi drivers do not? an empirical investigation of labor supply." *Journal of Labor Economics* **33(3)**: pp. 683–710.

Stanley, T. & H. Doucouliagos (2012): *Meta-Regression Analysis in Economics and Business*. London: Routledge.

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.

Stanley, T. D., H. Doucouliagos, & J. P. Ioannidis (2017): "Finding the Power to Reduce Publication Bias." *Statistics in Medicine* **36(10)**: pp. 1580–1598.

Stanley, T. D., S. B. Jarrell, & H. Doucouliagos (2010): "Could it be better to discard 90% of the data? A statistical paradox." *The American Statistician* **64(1)**: pp. 70–77.

Steel, M. F. J. (2020): "Model Averaging and its Use in Economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Stefansson, A. (2020): "Labor Supply Response to a Tax Holiday: The Take-Home from a Large and Salient Shock." *Working paper*, Uppsala University.

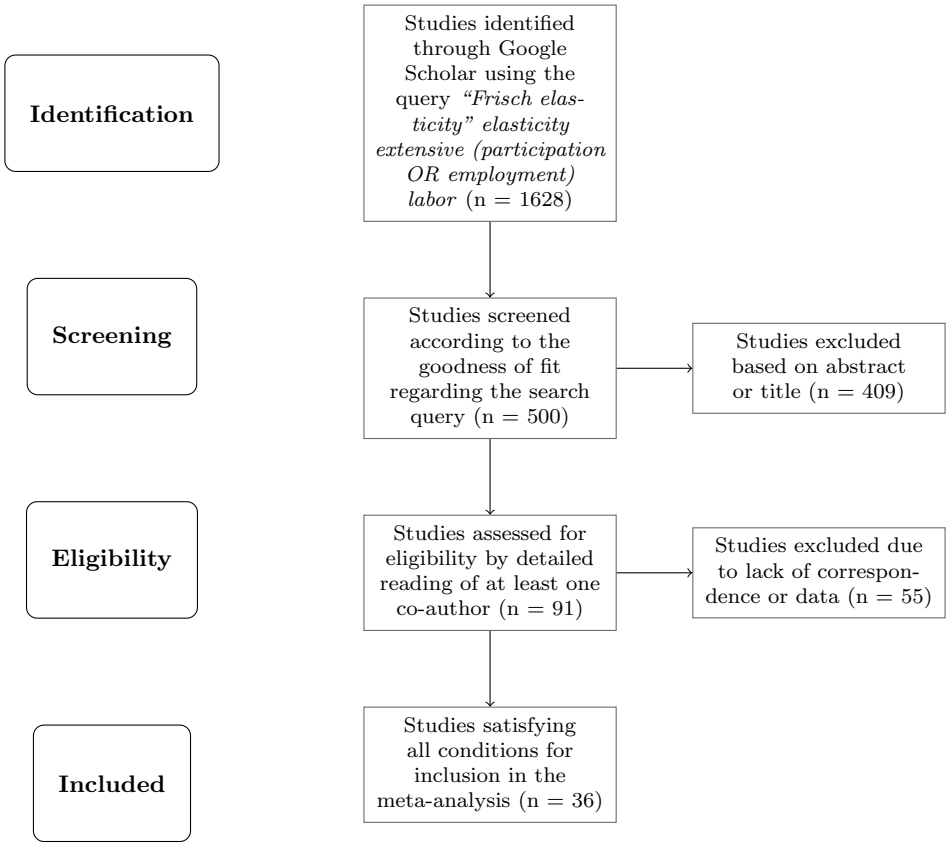
Whalen, C. & F. Reichling (2017): "Estimates of the Frisch Elasticity of Labor Supply: A Review." *Eastern Economic Journal* **43(1)**: pp. 37–42.

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.

Appendices (for online publication)

A Literature Search

Figure A1: The 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).

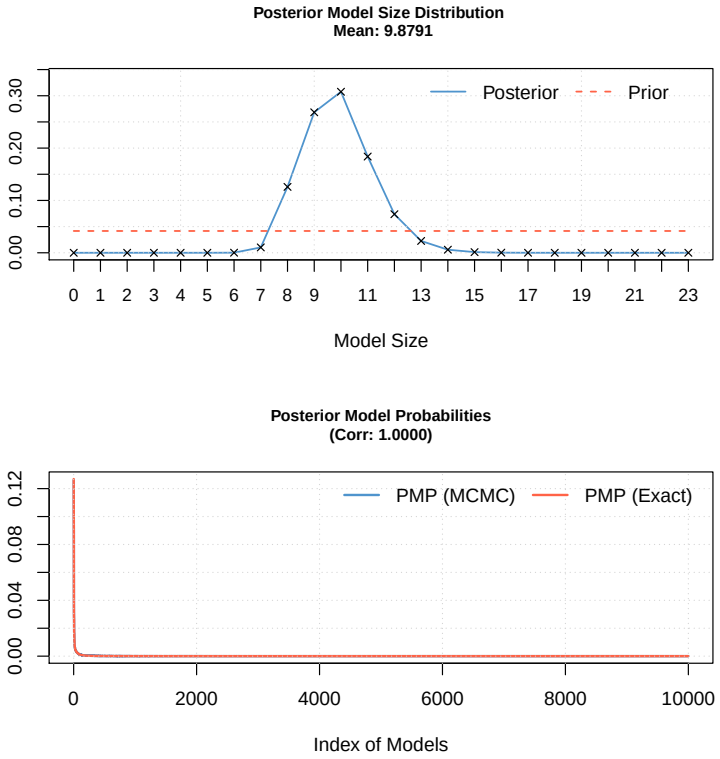
B Diagnostics and Robustness Checks

Table B1: Summary of the benchmark BMA estimation

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
9.8721	$3 \cdot 10^6$	$1 \cdot 10^6$	19.5064 mins	567,495
Modelspace	*Models visited*	*Topmodels*	*Corr PMP*	*No. Obs.*
8,388,608	6.8%	100	1.0000	723
Model prior	*g-prior*	*Shrinkage-stats*		
random / 11.5	UIP	Av=0.9986		

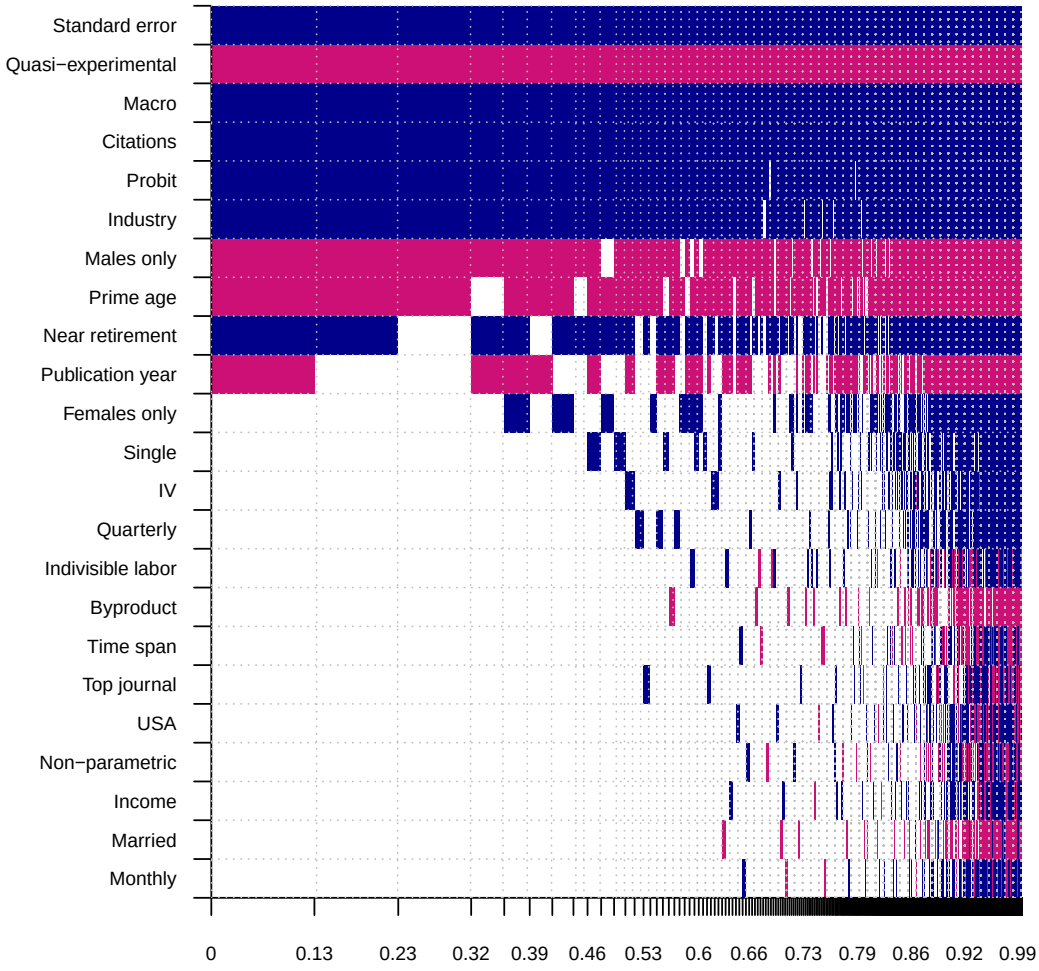
Notes: The results of this BMA specification are reported in Table 4. Based on Eicher *et al.* (2011) we employ unit information prior and, as suggested by George (2010), the dilution prior that takes into account potential collinearity.

Figure B1: Model size and convergence in the benchmark BMA model



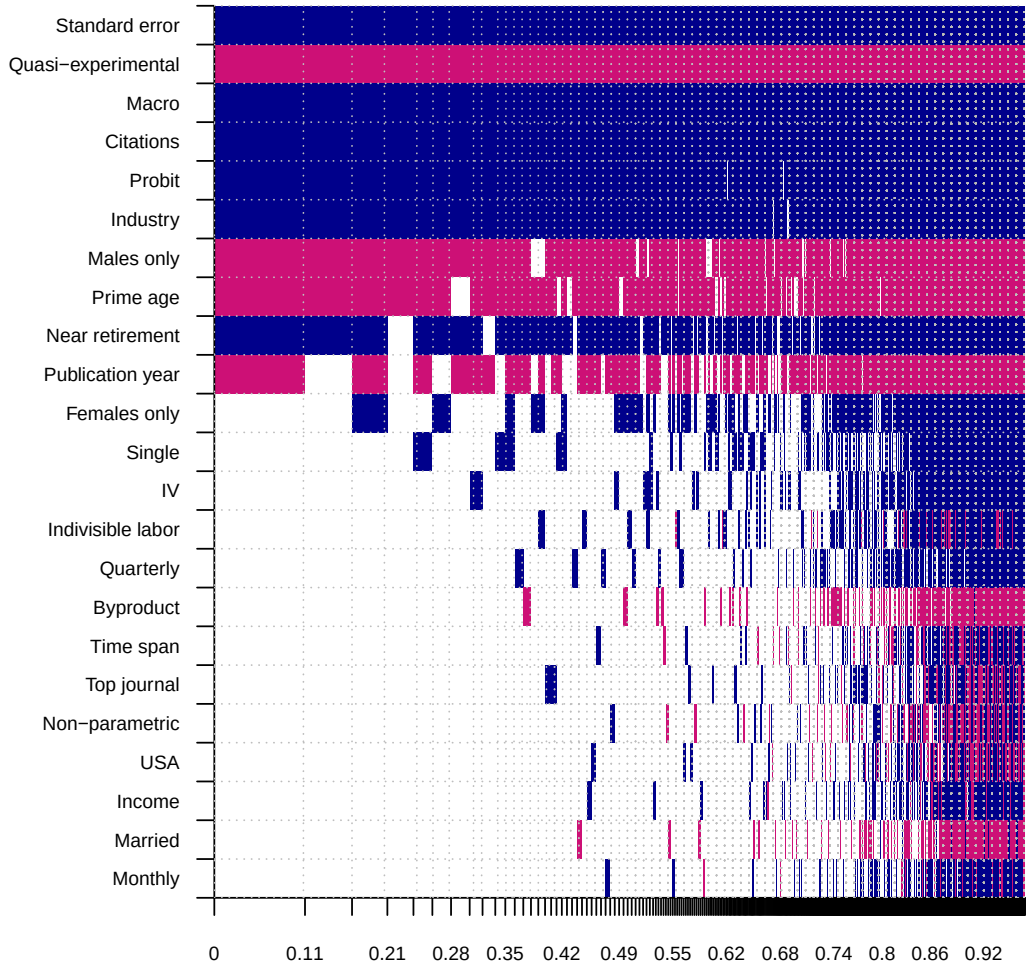
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table 4.

Figure B2: Model inclusion in BMA (BRIC g-prior)



Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior) and random model prior. Blue color (darker in grayscale) = the variable has a positive estimated sign. Red color (lighter in grayscale) = the variable has a negative estimated sign. No color = the variable is excluded from the given model. The numerical results are reported in Table B2.

Figure B3: Model inclusion in BMA (HQ g-prior)



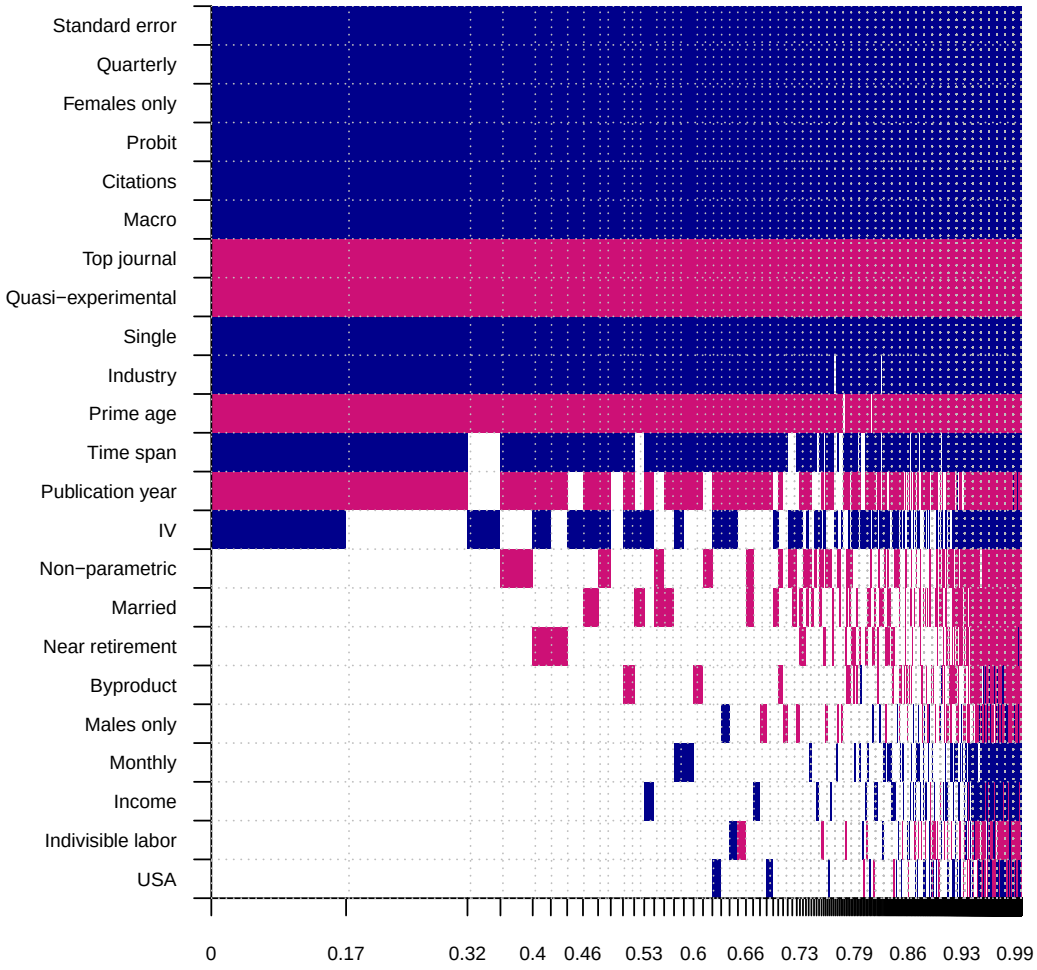
Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on HQ g-prior that asymptotically mimics the Hannan-Quinn criterion and random model prior. Blue color (darker in grayscale) = the variable has a positive estimated sign. Red color (lighter in grayscale) = the variable has a negative estimated sign. No color = the variable is excluded from the given model. The numerical results are reported in Table B2.

Table B2: Results of BMA with alternative priors

	BRIC g-prior			HQ g-prior		
Variable	Post. Mean	Post. SD	PIP	Post. Mean	Post. SD	PIP
Intercept	0.559	N.A.	1.000	0.673	N.A.	1.000
Standard error	1.331	0.146	1.000	1.297	0.147	1.000
Demographic characteristics						
Prime age	-0.092	0.048	0.863	-0.093	0.045	0.893
Near retirement	0.082	0.057	0.748	0.096	0.052	0.858
Females only	0.020	0.041	0.243	0.030	0.047	0.367
Males only	-0.102	0.044	0.909	-0.098	0.045	0.895
Married	-0.001	0.011	0.034	-0.001	0.014	0.058
Single	0.012	0.040	0.119	0.021	0.050	0.202
Income	0.001	0.008	0.037	0.001	0.010	0.065
Data characteristics						
Time span	0.000	0.006	0.045	0.000	0.008	0.079
Monthly	0.001	0.015	0.033	0.001	0.021	0.060
Quarterly	0.004	0.017	0.073	0.005	0.019	0.110
Industry	0.174	0.058	0.966	0.183	0.058	0.978
Macro	0.197	0.036	1.000	0.197	0.037	1.000
USA	0.001	0.011	0.043	0.000	0.012	0.066
Specification characteristics						
Indivisible labor	0.003	0.022	0.069	0.008	0.031	0.133
Quasi-experimental	-0.293	0.049	1.000	-0.306	0.048	1.000
Probit	0.244	0.068	0.988	0.232	0.067	0.989
Non-parametric	0.000	0.009	0.039	0.000	0.012	0.067
IV	0.005	0.022	0.078	0.009	0.030	0.141
Publication characteristics						
Publication year	-0.100	0.114	0.498	-0.133	0.116	0.654
Top journal	0.001	0.012	0.044	0.001	0.014	0.073
Citations	0.076	0.016	0.999	0.075	0.016	1.000
Byproduct	-0.003	0.018	0.050	-0.006	0.026	0.094
Observations	723			723		
Studies	36			36		

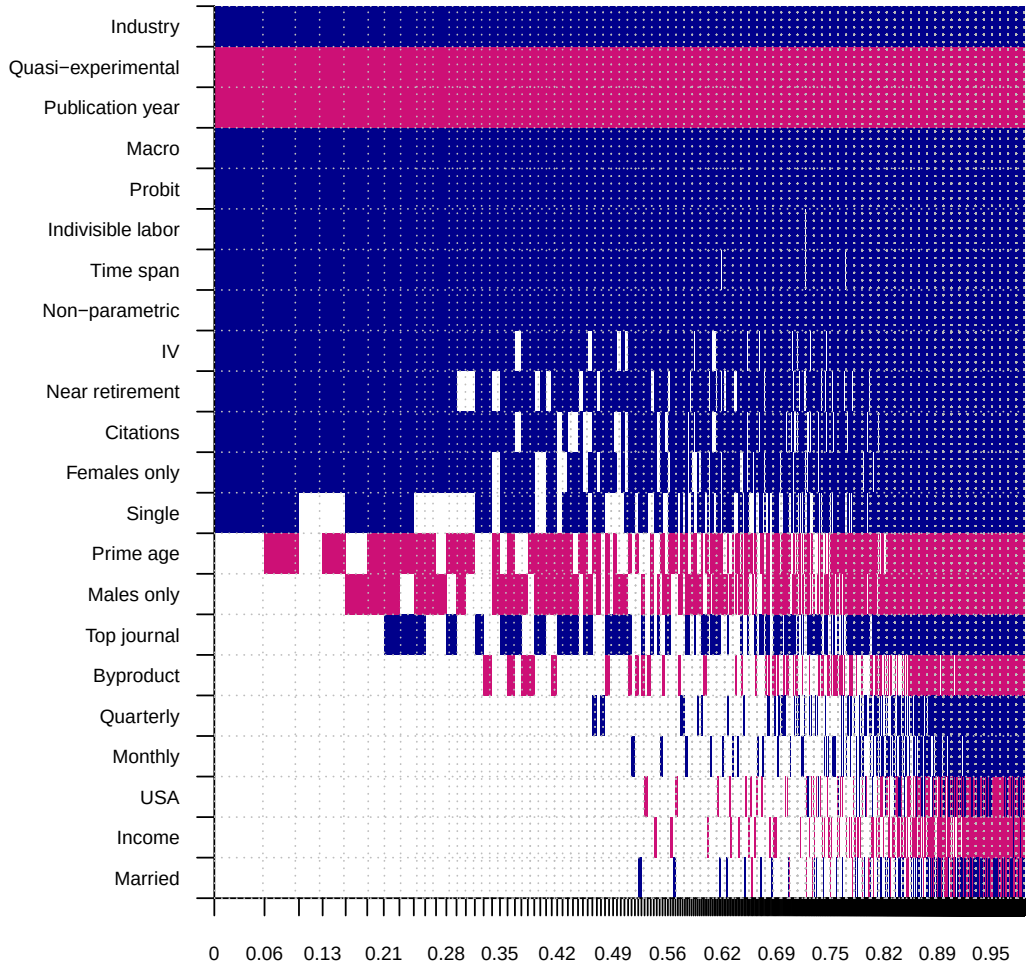
Notes: Response variable = The reported Frisch extensive elasticity, S.D. = standard deviation, PIP = Posterior inclusion probability. In the left panel, we apply BMA based on BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior). The right panel reports the results of BMA based on HQ g-prior, which asymptotically mimics the Hannan-Quinn criterion. Table 3 presents a detailed description of all variables.

Figure B4: Model inclusion in BMA (weighted)



Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on the unit information prior (UIP) recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010), which captures collinearity. The data used in BMA are weighted by the number of estimates per study. Blue color (darker in grayscale) = the variable has a positive estimated sign. Red color (lighter in grayscale) = the variable has a negative estimated sign. No color = the variable is excluded from the given model. Table B3 reports the numerical results.

Figure B5: Model inclusion in BMA excluding the standard error



Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on the unit information prior (UIP) recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010), which captures collinearity. In this setting, we exclude standard errors from the regressions. Blue color (darker in grayscale) = the variable has a positive estimated sign. Red color (lighter in grayscale) = the variable has a negative estimated sign. No color = the variable is excluded from the given model. Table B3 reports the numerical results.

Table B3: Alternative specifications of the baseline BMA model

	BMA (weighted)			BMA (without SE)		
Variable	Post. Mean	Post. SD	PIP	Post. Mean	Post. SD	PIP
Intercept	0.013	N.A.	1.000	2.250	N.A.	1.000
Standard error	1.615	0.117	1.000	N.A.	N.A.	N.A.
Demographic characteristics						
Prime age	-0.182	0.050	0.991	-0.053	0.053	0.589
Near retirement	-0.005	0.021	0.112	0.112	0.062	0.853
Females only	0.211	0.032	1.000	0.109	0.063	0.832
Males only	-0.002	0.013	0.081	-0.064	0.066	0.578
Married	-0.014	0.038	0.177	0.000	0.016	0.071
Single	0.266	0.062	0.998	0.094	0.090	0.602
Income	0.002	0.020	0.069	-0.001	0.011	0.076
Data characteristics						
Time span	0.064	0.033	0.882	0.091	0.027	0.987
Monthly	0.003	0.019	0.079	0.004	0.028	0.083
Quarterly	0.209	0.031	1.000	0.005	0.022	0.117
Industry	0.228	0.068	0.992	0.470	0.077	1.000
Macro	0.216	0.046	1.000	0.248	0.057	0.999
USA	0.000	0.008	0.064	0.000	0.016	0.082
Specification characteristics						
Indivisible labor	-0.001	0.014	0.067	0.246	0.070	0.991
Quasi-experimental	-0.196	0.044	1.000	-0.324	0.054	1.000
Probit	0.347	0.047	1.000	0.264	0.063	0.998
Non-parametric	-0.013	0.029	0.214	0.165	0.054	0.975
IV	0.054	0.059	0.535	0.182	0.084	0.904
Publication characteristics						
Publication year	-0.055	0.039	0.748	-0.679	0.101	1.000
Top journal	-0.164	0.036	1.000	0.050	0.067	0.439
Citations	0.087	0.016	1.000	0.050	0.028	0.842
Byproduct	-0.003	0.014	0.087	-0.019	0.050	0.186
Observations	723			723		
Studies	36			36		

Notes: Response variable = The Frisch extensive elasticity, S.D. = standard deviation, PIP = Posterior inclusion probability. In the left panel, variables are weighted by the inverse of the number of estimates per study. The right panel reports the results of BMA when standard errors are excluded. In both panels we employ BMA based on the UIP g-prior (Eicher *et al.*, 2011) and dilution model suggested by George (2010). Table 3 presents a detailed description of all variables.

Table B4: Results of frequentist model averaging

	Coeff.	S.E.	p-value
Intercept	1.203	0.316	0.000
Standard error	1.087	0.165	0.000
Demographic characteristics			
Prime age	-0.092	0.035	0.008
Near retirement	0.119	0.041	0.003
Females only	**0.101**	**0.039**	**0.009**
Males only	-0.060	0.039	0.122
Married	-0.003	0.056	0.961
Single	0.111	0.061	0.068
Income	0.023	0.038	0.549
Data characteristics			
Time span	0.028	0.027	0.292
Monthly	0.016	0.086	0.857
Quarterly	0.048	0.044	0.276
Industry	0.300	0.081	0.000
Macro	0.235	0.056	0.000
USA	-0.035	0.048	0.468
Specification characteristics			
Indivisible labor	0.112	0.059	0.057
Quasi-experimental	-0.332	0.062	0.000
Probit	0.212	0.069	0.002
Non-parametric	0.025	0.048	0.605
IV	0.099	0.058	0.087
Publication characteristics			
Publication year	**-0.329**	**0.098**	**0.001**
Top journal	-0.005	0.048	0.923
Citations	0.080	0.020	0.000
Byproduct	-0.099	0.068	0.147
Observations	723		
Studies	36		

Notes: We use Mallow's weights Hansen (2007) and the orthogonalization of the covariate space suggested byAmini & Parmeter (2012) to conduct the frequentist model averaging (FMA) exercise. The variables in bold are significant at the 5% level in FMA but not important in the benchmark BMA (with PIPs below 50%).