

Intertemporal Substitution in Labor Supply: A Meta-Analysis*

Ali Elminejad^a, Tomas Havranek^{a,b}, Roman Horvath^a, and Zuzana Irsova^a

^aCharles University, Prague

^bMeta-Research Innovation Center at Stanford

August 28, 2023

Abstract

The intertemporal substitution (Frisch) elasticity of labor supply governs the predictions of real business cycle models, New Keynesian models, and models of taxation. We show that the mean reported estimates, and consequently calibrations, are exaggerated due to publication bias. For both the intensive and extensive margins the literature provides over 700 estimates, with a mean of 0.5 in both cases. Correcting for publication bias and emphasizing quasi-experimental evidence reduces the mean intensive margin elasticity to 0.2 and renders the extensive margin elasticity tiny. An aggregate hours elasticity of about 0.25 is the most consistent with empirical evidence. To trace the differences in reported elasticities to differences in estimation context, we collect 23 variables reflecting study design and employ Bayesian and frequentist model averaging to address model uncertainty. On both margins the elasticity is systematically larger for women and workers near retirement, but not enough to support an aggregate hours elasticity above 0.5.

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

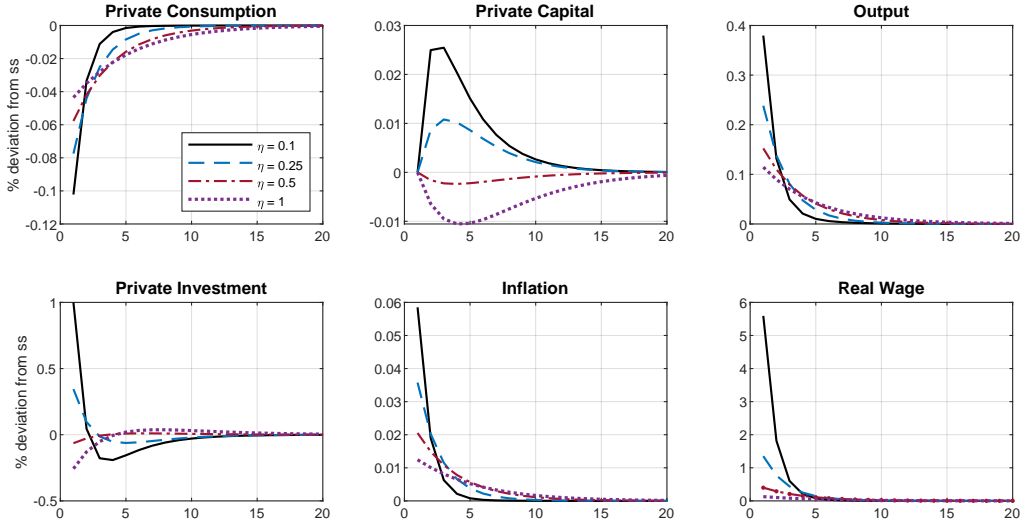
JEL Codes: C83, E24, J21

*An online appendix with data and code is available at meta-analysis.cz/frisch. Corresponding author: Zuzana Irsova, zuzana.irsova@ies-prague.org.

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 Gali (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 Gali (2015) and change the value of the Frisch elasticity (η) while leaving all the other parameters calibrated at the values used by Gali (2015).

For calibrations, researchers have increasingly relied on the entire corpus of microeconomic empirical literature instead of cherry-picking one or two preferred results. 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). While the CBO’s central estimate is conservative

and less than half the value of an earlier widely used survey of quasi-experimental evidence (Chetty *et al.*, 2013), which suggested overall aggregate hours Frisch elasticity of about 0.9, in this paper we show that even 0.4 is probably too large. The mean estimate reported in the literature is a systematically biased reflection of the underlying research results. For example, the Chetty *et al.* (2013) finding of a 0.9 overall elasticity matches our data remarkably well: the mean estimate in our dataset is 0.5 for both the intensive and extensive margins. Nevertheless, these summary statistics in our data are heavily distorted by publication bias and endogeneity in some studies. Conditional on the absence of publication bias and the availability of arguably exogenous time variation in wages, the literature is consistent with a tiny Frisch elasticity at the extensive margin (related to the decision whether to work) and an elasticity of 0.2 at the intensive margin (how much to work), consistent with about 0.25 for the overall elasticity.

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.

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

¹For recent papers on publication bias in economics, see Havranek (2015), Brodeur *et al.* (2016), Bruns & Ioannidis (2016), Ioannidis *et al.* (2017), Card *et al.* (2018), Christensen & Miguel (2018), Astakhov *et al.* (2019), DellaVigna *et al.* (2019), Bajzik *et al.* (2020), Blanco-Perez & Brodeur (2020), Brodeur *et al.* (2020), Fabo *et al.* (2021), Imai *et al.* (2021), Zigraiova *et al.* (2021), Gechert *et al.* (2022), Matousek *et al.* (2022), Ehrenbergerova *et al.* (2023), Havranek *et al.* (2023), and Yang *et al.* (2023). Earlier influential papers include Card & Krueger (1995), Ashenfelter *et al.* (1999), and Stanley (2001).

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.²

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

²Recently some authors have distinguished between publication bias (narrowly defined as the file-drawer problem) and p-hacking: see, for example, Brodeur *et al.* (2023); Irsova *et al.* (2023). When the distinction is made, publication bias denotes the decision not to publish the paper, while p-hacking denotes the effort to produce publishable results. Note that these two types of behavior are observationally equivalent in our data, so for parsimony we use the broader definition of publication bias, which also includes p-hacking. This broader definition of publication bias is common in most of the applied meta-analysis literature.

quantities, it follows that estimates and standard errors should not be correlated. The identification assumption can be violated in economics (for example, unobserved methods choices in primary studies may systematically affect both estimates and their standard errors),³ 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, 2023) 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, for the extensive margin, 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 extensive margin 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. 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 extensive margin Frisch elasticity. After correcting for both biases we find that the literature is consistent with a tiny elasticity. In contrast, the implied elasticity at the intensive margin is about 0.2.

³In addition, Keane & Neal (2023) show that for instrumental variable estimation, point estimates are likely to be correlated with standard errors.

The mean elasticity is important for the calibration of representative-agent models, but a small elasticity on average does not imply that workers do not 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 Mallows’ 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. Extensive margin 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. As the bottom line of our analysis, we use all the intensive and extensive margin 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 intensive margin elasticity is around 0.2, while the elasticity is tiny for the extensive margin. A value of 0.25 for the overall elasticity is most consistent with the literature. The overall elasticities corresponding to women and workers near retirement are around 0.3–0.4.

Two 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. Given the focus on 6 estimates, Chetty *et al.* (2013) cannot examine publication bias. Second, Martinez *et al.* (2021) use the natural experiment of tax holidays in Switzerland to estimate the Frisch elasticity. Given their 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. Our results are similar qualitatively to Martinez *et al.* (2021): substitution is negligible at the extensive margin and small at the intensive margin. Quantitatively, though, Martinez *et al.* (2021) find an overall elasticity of 0.025, while our estimate is an order of magnitude larger, about 0.25. Both results are hard to reconcile with common calibrations of macroeconomic models.

A qualification is in order regarding the object under examination in the empirical literature on the Frisch elasticity. Conceptually, the elasticity represents the preferences of households. But researchers, even when blessed with high-quality quasi-experimental data, observe labor market outcomes that are also affected by salience and frictions (Chetty *et al.*, 2009; Chetty, 2012; Sigurdsson, 2023b). It may be that workers have relatively elastic labor supply preferences but do not change their behavior because they are not sufficiently attentive to the change in net wages or because they face substantial adjustment costs, search frictions, or liquidity constraints. The literature does not provide enough information to allow us to disentangle the correct Frisch elasticity from the confounding effects of salience and frictions. Conceptually, this is an important limitation of our analysis (and the empirical literature on labor supply elasticities). In practice, however, the reduced-form elasticities that we survey are informative regarding the real-world behavior of households with respect to temporary changes in wages, and hence our results are still useful for calibration.

This paper includes two meta-analyses: one for the extensive margin, the other for the intensive margin. Because these are economically distinct concepts, they cannot be reasonably pooled together in one meta-analysis. To avoid duplicating meta-analysis outputs and discussion, in the main text we focus on the extensive margin, for which quasi-experimental evidence is more abundant; the meta-analysis of intensive margin elasticities is available in Appendix A. The meta-analysis methods in both parts are identical, and any substantial differences in results are discussed in the main text. Subsection 4.4 in the main body of the paper summarizes the results of both meta-analyses.

2 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 B 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 since working papers are intended for eventual publication and any mechanisms that lead to preference for positive or significant estimates in journal articles also apply to working papers, as shown, for example, by Rusnak *et al.* (2013). We terminate the search on July 1, 2023, and do not add any studies beyond that date. The final sample includes 709 intensive margin estimates from 40 studies (see Appendix A for a meta-analysis of intensive margin elasticities) and 762 extensive margin estimates from 38 studies (Table 1) covering a quarter century of research on labor supply elasticities. Appendix C provides details on how the elasticities are estimated, and Table B1 shows how we collected estimates from individual papers. The data are available in an online appendix at meta-analysis.cz/frisch.

Figure 2 shows the distribution of Frisch elasticities at the extensive margin reported in the literature. The mean (0.48) is substantially larger than the median (0.35), but overall the literature appears to be quite consistent with the CBO’s calibration at 0.4 (which, however, takes into account both the intensive and extensive margins). We also observe that the economically

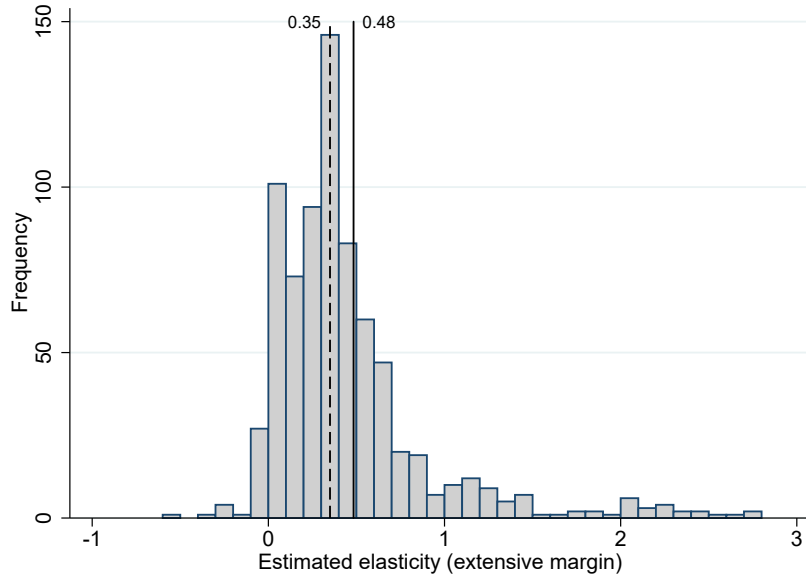
Table 1: Studies included in the meta-analysis of extensive margin elasticities

Attanasio <i>et al.</i> (2018)	Haan & Uhlenhorff (2013)
Beffy <i>et al.</i> (2019)	Inoue (2015)
Bianchi <i>et al.</i> (2001)	Karabarbounis (2016)
Blundell <i>et al.</i> (2016a)	Keane & Wasi (2016)
Blundell <i>et al.</i> (2016b)	Kimmel & Kniesner (1998)
Borella <i>et al.</i> (2023)	Kneip <i>et al.</i> (2019)
Brown (2013)	Kuroda & Yamamoto (2008)
Caldwell (2019)	Looney & Singhal (2006)
Card & Hyslop (2005)	Manoli & Weber (2011)
Carrington (1996)	Manoli & Weber (2016)
Chang & Kim (2006)	Martinez <i>et al.</i> (2021)
Chang <i>et al.</i> (2019)	Mustre-del Rio (2011)
Erosa <i>et al.</i> (2016)	Mustre-del Rio (2015)
Espino <i>et al.</i> (2017)	Oettinger (1999)
Fiorito & Zanella (2012)	Ong (2019)
French & Stafford (2017)	Park (2020)
Gine <i>et al.</i> (2017)	Peterman (2016)
Gourio & Noual (2009)	Sigurdsson (2023a)
Gruber & Wise (1999)	Stafford (2015)

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. 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, industry-specific data, and use of wage ratios. We also include dummy variables reflecting econometric techniques (e.g., probit, 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

Figure 2: Estimates are most commonly around 0.4

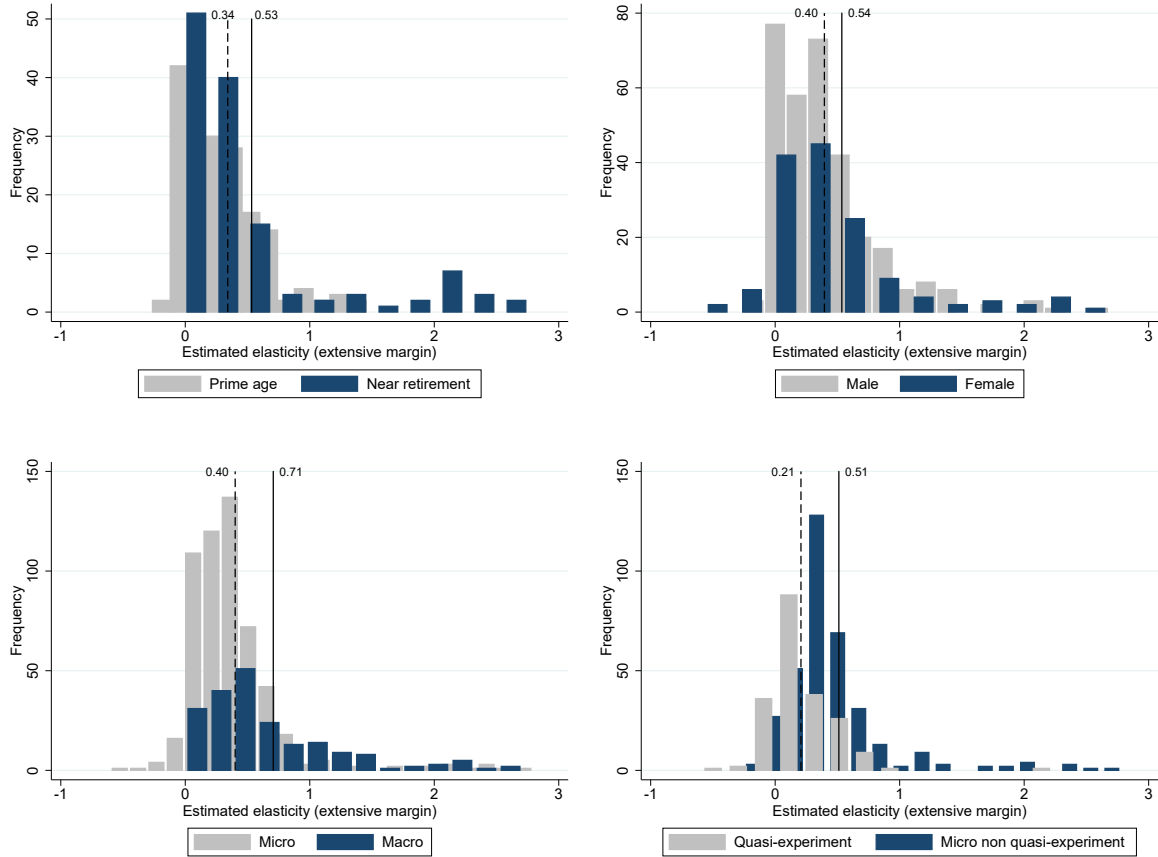


Notes: The solid line denotes the sample mean (0.48); the dashed line denotes the sample median (0.35). Estimates smaller than -1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

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 4.

An important variable for meta-analysis is the standard error of the reported estimate. Nevertheless, for some 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 0.14 for gender and 0.19 for age. A larger difference arises between estimates using micro (0.40 on average) and macro data (0.71). Note that we consider 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 is a substantial difference between micro estimates based on quasi-experimental data (0.21 on average) and

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.

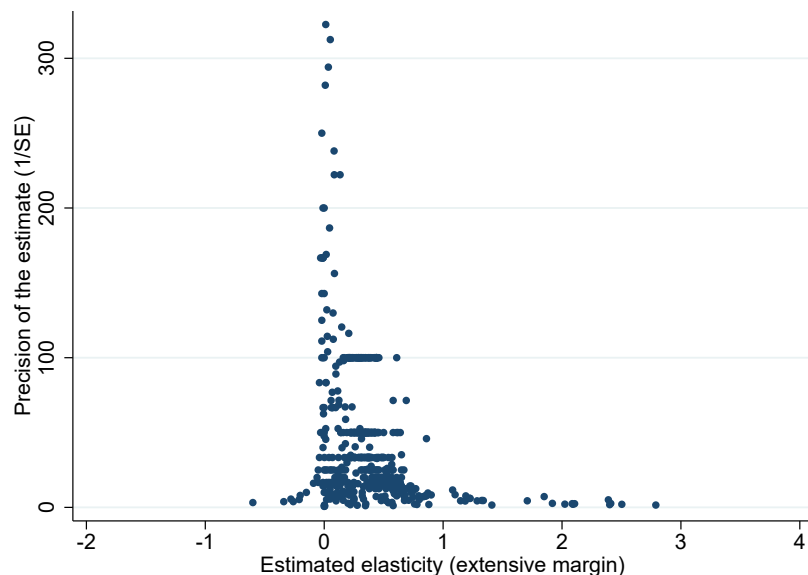
non-quasi-experimental data, which use variation in taxes or wages in the absence of significant policy shifts (0.51). These stylized facts suggest both genuine differences in the elasticity (which are however modest) and the importance of proper identification. Studies more likely to exploit truly exogenous 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.

3 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.

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.

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 0.25.) 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 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:

$$\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}, \quad (1)$$

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	MAIVE
Publication bias (<i>Standard error</i>)	1.689 ^{***} (0.264) [1.05, 2.36]	0.887 ^{***} (0.271) -	2.592 ^{***} (0.530) [1.55, 3.86]	2.173 ^{***} (0.227) [1.68, 2.70]	3.056 ^{**} (1.500) {0.53, 6.47}
Effect beyond bias (<i>Constant</i>)	0.288 ^{***} (0.0442) [0.11, 0.37]	0.356 ^{***} (0.0252) -	0.211 ^{***} (0.0441) [0.06, 0.29]	0.243 ^{***} (0.0470) [0.15, 0.34]	0.350 ^{***} (0.0463) {0.06, 0.74}
First stage F-stat					31.2
Observations	762	762	762	762	603
Studies	38	38	38	38	23
Panel B: Nonlinear tests					
	Ioannidis <i>et al.</i> (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2023)
Effect beyond bias	0.208 ^{***} (0.055)	0.354 ^{***} (0.064)	0.142 ^{***} (0.009)	0.063 (0.077)	0.365 ^{***} (0.092)
Observations	762	762	762	762	762
Studies	38	38	38	38	38

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. MAIVE = meta-analysis instrumental variable estimator (Irsova *et al.*, 2023); 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. For MAIVE, in curly brackets we show the weak-instrument-robust Anderson-Rubin 95% confidence interval. Panel B presents the mean elasticity corrected for publication bias using nonlinear techniques described in the main text. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panel A of Table 2 presents the results of estimating (1). 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 (38 studies) we additionally report confidence intervals based on wild bootstrap where applicable. (For the instrumental variable estimator described later we instead report the weak-instrument-robust Anderson-Rubin confidence interval.) 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.

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 31). 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. The point estimates of the slope coefficient range from 0.9 (fixed effects) to 3.1 (instrumental variables). Confidence intervals based on wild bootstrap range from 1 to 4, and the median estimate is 2.2. Three out of the five techniques suggest a slope coefficient above 2. Overall, it seems that 2 is a relatively conservative estimate for the slope coefficient, which translates to strong publication bias. To see this, consider a hypothetical case in which the true elasticity was zero. Then the true mean t-statistic should be zero as well. But a slope coefficient of 2 in meta-regression is consistent with a mean reported t-statistic of 2 since in such a case point estimates are on average twice the standard error. So a slope of 2 would suggest a positive and significant *reported* effect on average even in the absence of an *underlying* effect: a dramatic change in inference due to publication bias. 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

0.21 (precision-weighted specification) to 0.36 (fixed effects) with a median estimate of 0.29 and bootstrapped confidence intervals from 0.1 to 0.4. These results imply that publication bias exaggerates the mean elasticity almost twofold.

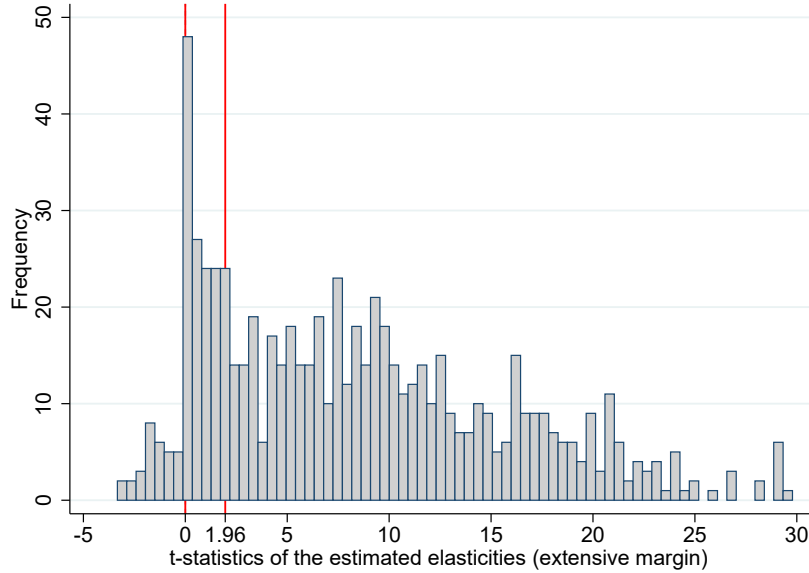
A problem of the funnel asymmetry test we have not yet addressed is the assumption that publication bias is a linear function of the standard error. The assumption is tenuous for small standard errors if the underlying elasticity is not zero. Consider, for example, the case when the true Frisch elasticity at the extensive margin is 0.29. 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 0.14 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 (2023) 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 show the values of t-statistics associated with changing the sign and achieving statistical significance at the 5% level, respectively.

The results of the nonlinear techniques are similar to the results reported previously for the funnel asymmetry tests but suggest an even smaller corrected mean elasticity. In all cases the mean corrected for publication bias is smaller than the simple reported mean of 0.49: estimates range from 0.06 (Furukawa, 2021) to 0.37 (van Aert & van Assen, 2023). The median estimate for the nonlinear techniques is 0.21, compared to the 0.29 value in the previous panel: together, the two panels suggest that 0.25 is a reasonable estimate for the mean Frisch elasticity at the extensive margin. We conclude that publication bias in the literature is substantial and likely to exaggerate the mean reported elasticity approximately twofold. Appendix A shows that the findings are similar for intensive margin elasticities, implying only slightly smaller publication bias. 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.* (2022), not reported here, confirm the observation.

In our baseline analysis we pool together structural and quasi-experimental estimates of the elasticity. Table D1, presented in Appendix D, shows the analysis of publication bias separately for the subsample of quasi-experimental estimates. The implied values for the corrected mean Frisch elasticity are smaller than in the entire sample, around 0.15. Note that quasi-experimental data are often examined for demographic groups (women, workers near retirement) that are likely to display a larger elasticity than the population as a whole; in the next section we will derive an estimate conditional on quasi-experimental data for the general population. Regarding structural estimates of the elasticity, Keane & Neal (2023) show that with instrumental variables, point estimates are correlated with standard errors, and the correlation depends on instrument strength. We find some tentative evidence that the correlation may be stronger with weaker instruments (Table D2). A mechanical correlation between estimates and standard errors is a grave problem for almost all meta-analysis methods. As we have noted, two of our techniques allow for such a correlation in the absence of publication bias. First, the MAIVE approach due to Irsova *et al.* (2023), in which a function of sample size is used as an instrument for the standard error. But MAIVE may not fully address the problem because sample size is related to instrument strength. The p-uniform* approach, described and reported earlier, is a more promising remedy in this case since it relies on identification unrelated to the correlation between estimates and standard errors.

4 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.

4.1 Variables

Demographic characteristics A potentially important source of heterogeneity stems from the demographic characteristics of the samples used in primary studies. We define six 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 (38%) 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 19% 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. Finally, two dummy variables control for the marital status of the people examined. Only 5% of estimates correspond to married workers only, and 4% 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

Table 3: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
Frisch elasticity	The estimated extensive margin Frisch elasticity (response variable).	0.48	0.63
Standard error	The standard error of the estimate.	0.10	0.17
<i>Demographic characteristics</i>			
Prime age	= 1 if the sample only consists of people between 25 and 55 years of age.	0.21	0.41
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.	0.19	0.39
Males only	= 1 if the sample consists of males only.	0.42	0.49
Married	= 1 if the sample consists of married people only.	0.05	0.23
Single	= 1 if the sample consists of single people only.	0.04	0.20
<i>Data characteristics</i>			
Time span	The logarithm of the data time span used to estimate the elasticity.	2.23	0.88
Monthly	= 1 if the data frequency is monthly (reference category: annual).	0.02	0.14
Quarterly	= 1 if the data frequency is quarterly (reference category: annual).	0.23	0.42
Ratio	= 1 if a wage ratio (income divided by hours) is used to estimate the elasticity, =0 if direct wage measures are used.	0.71	0.45
Industry	= 1 if the sample consists of workers in a specific industry (reference category: whole economy data).	0.11	0.32
Macro	= 1 if the estimate uses aggregated data (reference category: micro).	0.27	0.44
USA	= 1 if the estimate uses data for the US.	0.66	0.47
<i>Specification characteristics</i>			
Indivisible labor	= 1 if the labor supply is assumed to be indivisible in the estimation framework.	0.33	0.47
Quasi-experimental	= 1 if the estimation framework uses quasi-experimental identification.	0.27	0.44
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).	0.18	0.38
<i>Publication characteristics</i>			
Publication year	The logarithm of the publication year the study.	3.47	0.20
Top journal	= 1 if the estimate is published in a top five journal in economics.	0.25	0.44
Citations	The logarithm of the number of per-year citations of the study in Google Scholar.	1.52	1.31
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.	0.07	0.25

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

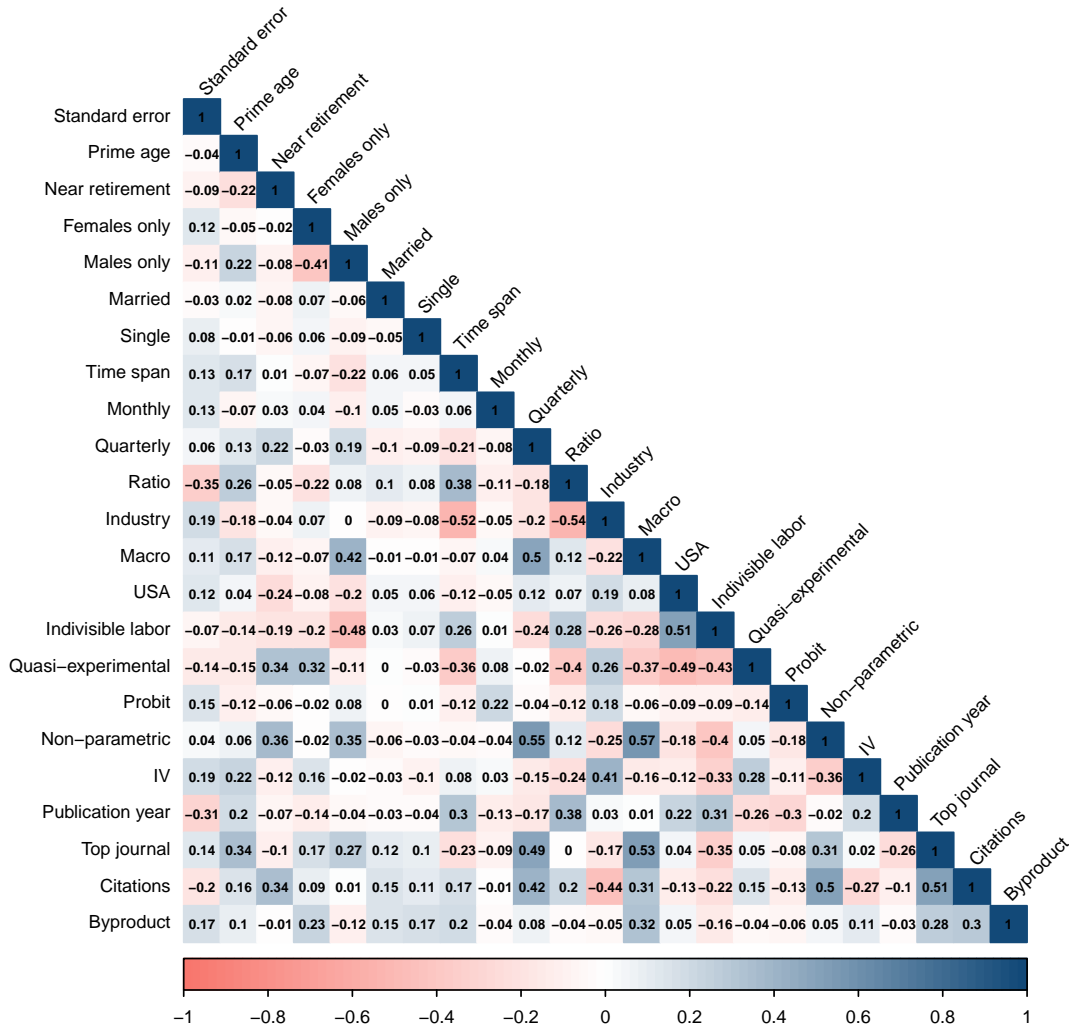
frequency is the relevant time frame for business cycle analysis. In addition, we control for the fact whether a wage ratio (income divided by hours) is used to estimate the elasticity; Keane (2011) notes that such an approach can contribute to attenuation bias. The dummy variable “Industry” controls for the fact whether the estimate uses data from a specific industry. About 66% 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 (73%) 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 0.21, 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 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 40% of estimates. Probit models are used only in 5% of estimates, while the instrumental variables and non-parametric methods are used in 18% 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: Correlations among explanatory variables are modest



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

Figure 6 shows that correlations among the variables are not extensive. The largest correlation coefficient is 0.57, 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 for women and workers near retirement, macro studies often use data at the quarterly frequency, time spans used in studies have been increasing recently, and studies published in top journals tend to be frequently cited.

4.2 Estimation

The intuitive approach to model heterogeneity is to regress the reported elasticities on all the variables introduced above. But that approach is incorrect because it ignores model uncertainty: while we 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). BMA has been used in meta-analysis, for example, by Havranek & Irsova (2017); Havranek *et al.* (2017, 2018a,b).

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

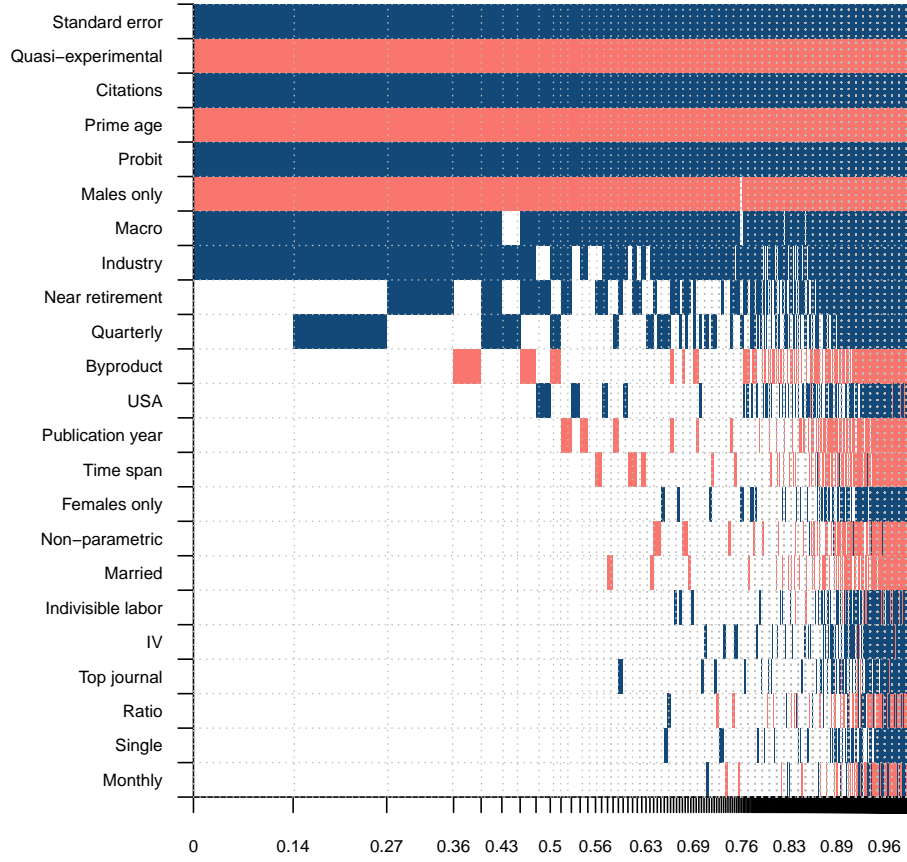
the dilution prior suggested by George (2010), which takes into account the collinearity of the variables included in each model. The prior multiplies the model probabilities by the determinant of the correlation matrix of the variables. Higher collinearity means that the determinant is closer to zero, 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.

4.3 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 8 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, four variables have PIP equal to 1, indicating that they are *decisive* variables (in the classification of Raftery *et al.* 1997); two 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. 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 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.

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.

Table 4: Why do estimates of the elasticity vary?

Response variable: Frisch elasticity (extensive margin)	Bayesian model averaging (baseline model)			Ordinary least squares (frequentist check)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Intercept	0.325	NA	1.000	0.289	0.025	0.000
Standard error	1.381	0.120	1.000	1.384	0.120	0.000
<i>Demographic characteristics</i>						
Prime age	-0.150	0.030	1.000	-0.156	0.045	0.001
Near retirement	0.034	0.047	0.390			
Females only	0.003	0.014	0.057			
Males only	-0.113	0.032	0.980	-0.116	0.049	0.023
Married	-0.002	0.015	0.047			
Single	0.001	0.012	0.034			
<i>Data characteristics</i>						
Time span	-0.002	0.010	0.073			
Monthly	0.000	0.014	0.029			
Quarterly	0.030	0.045	0.363			
Ratio	0.000	0.008	0.035			
Industry	0.128	0.066	0.859	0.146	0.062	0.024
Macro	0.134	0.051	0.942	0.145	0.052	0.009
USA	0.007	0.024	0.112			
<i>Specification characteristics</i>						
Indivisible labor	0.001	0.013	0.043			
Quasi-experimental	-0.285	0.042	1.000	-0.279	0.033	0.000
Probit	0.232	0.057	0.995	0.233	0.099	0.024
Non-parametric	-0.002	0.014	0.055			
IV	0.001	0.012	0.042			
<i>Publication characteristics</i>						
Publication year	-0.010	0.038	0.087			
Top journal	0.001	0.010	0.039			
Citations	0.067	0.013	1.000	0.074	0.014	0.000
Byproduct	-0.016	0.042	0.165			
Observations	762			762		
Studies	38			38		

Notes: The response variable is the Frisch elasticity of labor supply at the extensive margin. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = Posterior inclusion probability, SE = 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.

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 prime age workers are systematically smaller than elasticities for other age groups, especially workers near retirement. The findings confirm the patterns in the literature shown earlier in Figure 2 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.

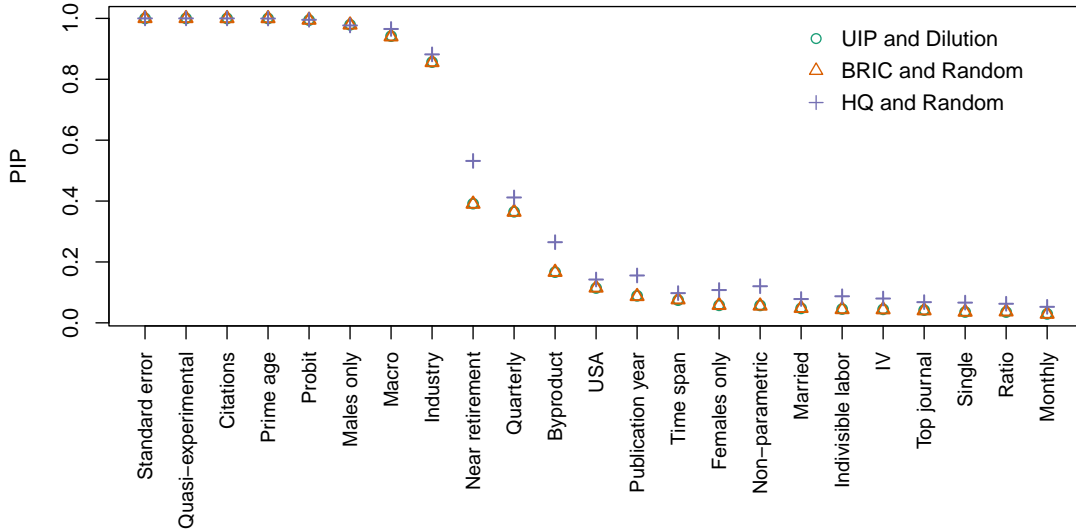
Data characteristics. Our results indicate no systematic effect of the time span, data frequency, and definition of wages used in the primary study on the reported elasticity. We do not find evidence that the US-based estimates are systematically different from estimates reported for other countries. In contrast, elasticities obtained from macro data tend to be 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 2.

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 concentrates 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.

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. PIP = posterior inclusion probability.

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 D. Finally, we apply frequentist model averaging (FMA), which does not need priors. We use Mallows’s weights (Hansen, 2007) and the orthogonalization of covariate space suggested by Amini & Parmeter (2012). Table D4 reports the results of FMA exercise. The robustness checks corroborate our main results. Regarding the analysis of heterogeneity in intensive margin elasticities examined in Appendix A, the results are similar to those for the extensive margin in several respects: publication bias is still important and elasticities are larger for women and workers near retirement. In contrast, for the intensive margin quasi-experimental identification brings larger estimated elasticities compared to other approaches that rely on micro data.

4.4 Implied Elasticities

As the bottom line of our analysis we compute the Frisch elasticity, both on the intensive and extensive margins, 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 a large, recent, and well-published quasi-experimental study, Martinez *et al.* (2021). 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.

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 relatively well in the case of the Frisch elasticity and yields results that are slightly more conservative (that is, correct for publication bias less aggressively) than nonlinear techniques. 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. Because of measurement error considerations, we prefer when direct wage measures are used, not wage ratios. 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 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 panel presents our subjective best practice defined in the previous paragraph. In the second panel we conduct a similar exercise but instead of selecting aspects of best practice subjectively we choose the aspects of the baseline estimation in Martinez *et al.* (2021). To avoid false precision, for practical purposes we prefer to round the results. The mean intensive margin elasticity is around 0.2 in both panels. The extensive margin elasticity is very small but not really zero. So, for the aggregate hours elasticity in a representative agent model, 0.25 seems to be the value most consistent with the empirical literature after correction for biases. The elasticities are larger for some demographic groups: especially women and workers near retirement. For these subgroups calibrations of the aggregate hours elasticity up to 0.4 can be backed by the literature. (For completeness, the table also includes elasticities for single and married workers, although these results should be interpreted with caution because only a small fraction of the estimates in our sample correspond to these subgroups.) Note also the wide confidence intervals: while our results do not explicitly support

Table 5: Mean elasticities implied by the literature

Panel A: Subjective best practice	Extensive margin		Intensive margin	
	Mean	95% CI	Mean	95% CI
Overall	0.03	[-0.24, 0.30]	0.24	[-0.03, 0.52]
Near retirement	0.14	[-0.10, 0.39]	0.25	[-0.09, 0.59]
Prime age	-0.09	[-0.42, 0.23]	0.15	[-0.09, 0.40]
Women	0.12	[-0.10, 0.35]	0.27	[-0.01, 0.55]
Married women	0.10	[-0.12, 0.33]	0.31	[0.03, 0.59]
Single women	0.19	[-0.04, 0.42]	0.12	[-0.16, 0.40]
Men	-0.02	[-0.34, 0.29]	0.17	[-0.09, 0.43]
Married men	-0.04	[-0.35, 0.27]	0.21	[-0.05, 0.48]
Single men	0.05	[-0.25, 0.34]	0.02	[-0.24, 0.28]
Panel B: Martinez <i>et al.</i> (2021)	Extensive margin		Intensive margin	
	Mean	95% CI	Mean	95% CI
Overall	0.02	[-0.11, 0.16]	0.18	[-0.08, 0.44]
Near retirement	0.13	[-0.08, 0.34]	0.19	[-0.15, 0.54]
Prime age	-0.11	[-0.22, 0.01]	0.09	[-0.12, 0.31]
Women	0.11	[-0.02, 0.24]	0.21	[-0.07, 0.49]
Married women	0.09	[-0.04, 0.23]	0.25	[0.01, 0.50]
Single women	0.18	[0.02, 0.35]	0.06	[-0.19, 0.31]
Men	-0.04	[-0.21, 0.14]	0.11	[-0.17, 0.39]
Married men	-0.05	[-0.23, 0.12]	0.15	[-0.10, 0.41]
Single men	0.04	[-0.13, 0.21]	-0.04	[-0.32, 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 rows show estimates for different demographic groups. In the first panel 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 lower panel we do not define best practice ourselves but use the characteristics used by Martinez *et al.* (2021). The 95% confidence intervals are reported in parentheses. The results for single men and women should be interpreted with caution because these subgroups are examined by a small fraction of the literature (around 2% of the estimates on average).

calibrations above 0.5, elasticities slightly above this value cannot be ruled out. Although our central estimate of roughly 0.25 is below the lower bound of the range of elasticities used for the calibration of the CBO's model mentioned in the Introduction, the CBO's central estimate (0.4) can be consistent with the literature. Nevertheless, the literature is clearly inconsistent with aggregate hours elasticities above 0.8 commonly used in macro models.

5 Conclusion

A general implication of our results is that it is risky to calibrate a parameter of a structural model, or indeed any policy, based on the mean estimate of that parameter reported in the literature. The reported mean is often a biased 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 for both the intensive and extensive margin elasticities. 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. 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 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 qualitatively consistent with our large meta-analysis: intertemporal substitution in labor supply is weak.

If a high-quality primary study can serve as a good guide for calibration or policy, why bother with a meta-analysis? Publication bias is not a problem of literature surveys exclusively—it can affect the results reported in any primary study. 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 small underlying effect, it would be 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.

An important problem we cannot fully address is potential attenuation bias, the “iron law of econometrics” (Hausman, 2001). Wages 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 understate the strength of intertemporal substitution. A crude way how to evaluate the extent of (classical) 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 net wages, 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 addition, elasticities derived from wage ratios tend to be similar to elasticities derived from direct wage measures. Although we fail to find evidence of substantial attenuation bias, we cannot rule it out.

References

- AARONSON, D. & E. FRENCH (2009): “The effects of progressive taxation on labor supply when hours and wages are jointly determined.” *Journal of Human Resources* **44**(2): pp. 386–408.
- VAN AERT, R. C. & M. VAN ASSEN (2023): “Correcting for publication bias in a meta-analysis with the p-uniform* method.” *Working paper*, Tilburg University & Utrecht University.
- ALTONJI, J. G. (1986): “Intertemporal substitution in labor supply: Evidence from micro data.” *Journal of Political Economy* **94**(3, Part 2): pp. S176–S215.
- AMINI, S. M. & C. F. PARMETER (2012): “Comparison of model averaging techniques: Assessing growth determinants.” *Journal of Applied Econometrics* **27**(5): pp. 870–876.
- ANDREWS, I. & M. KASY (2019): “Identification of and correction for publication bias.” *American Economic Review* **109**(8): pp. 2766–2794.
- ANGRIST, J. D. (1991): “Grouped-data estimation and testing in simple labor-supply models.” *Journal of Econometrics* **47**(2-3): pp. 243–266.
- ANGRIST, J. D., S. CALDWELL, & J. V. HALL (2021): “Uber versus taxi: A driver’s eye view.” *American Economic Journal: Applied Economics* **13**(3): pp. 272–308.
- 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.
- ASTAKHOV, A., T. HAVRANEK, & J. NOVAK (2019): “Firm size and stock returns: A quantitative survey.” *Journal of Economic Surveys* **33**(5): pp. 1463–1492.
- 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.
- BAJZIK, J., T. HAVRANEK, Z. IRSOVA, & J. SCHWARZ (2020): “Estimating the Armington elasticity: The importance of study design and publication bias.” *Journal of International Economics* **127**(C).
- BATTISTI, M., R. MICHAELS, & C. PARK (2023): “Labor supply within the firm.” *Journal of Labor Economics* (forthcoming).
- BEFFY, M., R. BLUNDELL, A. BOZIO, G. LAROQUE, & M. TO (2019): “Labour supply and taxation with restricted choices.” *Journal of Econometrics* **211**(1): pp. 16–46.
- 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." *The 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.
- BORELLA, M., M. DE NARDI, & F. YANG (2023): "Are marriage-related taxes and social security benefits holding back female labour supply?" *The Review of Economic Studies* **90(1)**: pp. 102–131.
- BREDEMEIER, C., J. GRAVERT, & F. JUESSEN (2019): "Estimating labor supply elasticities with joint borrowing constraints of couples." *Journal of Labor Economics* **37(4)**: pp. 1215–1265.
- BRODEUR, A., S. CARRELL, D. FIGLIO, & L. LUSHER (2023): "Unpacking p-hacking and publication bias." *American Economic Review* **forthcoming**.
- 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(C)**: 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. & E. OEHLSEN (2022): "Gender differences in labor supply: Experimental evidence from the gig economy." *Working paper*, University of California, Berkeley Working Paper.
- 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. JOHANNESSEN, 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(9)**: 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 (2011): "Interpreting labor supply regressions in a model of full-and part-time work." *American Economic Review* **101(3)**: pp. 476–481.
- 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. (2012): "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* **80(3)**: pp. 969–1018.
- 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.
- CHETTY, R., W. LOONEY, & K. KROFT (2009):

- “Salience and taxation: Theory and evidence.” *American Economic Review* **99**(4): pp. 1145–77.
- 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.
- DOMELJ, D. & M. FLODEN (2006): “The labor-supply elasticity and borrowing constraints: Why estimates are biased.” *Review of Economic Dynamics* **9**(2): pp. 242–262.
- 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.
- EHRENBERGEROVA, D., J. BAJZIK, & T. HAVRANEK (2023): “When Does Monetary Policy Sway House Prices? A Meta-Analysis.” *IMF Economic Review* **71**(2): pp. 538–573.
- 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 (2022): “Detecting p-hacking.” *Econometrica* **90**(2): pp. 887–906.
- 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.
- FABO, B., M. JANCOKOVA, E. KEMPF, & L. PASTOR (2021): “Fifty shades of QE: Comparing findings of central bankers and academics.” *Journal of Monetary Economics* **120**(C): pp. 1–20.
- FARBER, H. S. (2015): “Why you can’t find a taxi in the rain and other labor supply lessons from cab drivers.” *The Quarterly Journal of Economics* **130**(4): pp. 1975–2026.
- 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, E. (2005): “The effects of health, wealth, and wages on labour supply and retirement behaviour.” *The Review of Economic Studies* **72**(2): pp. 395–427.
- 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.
- GALI, J. (2015): *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, second edition.
- GECHERT, S., T. HAVRANEK, Z. IRSOVA, & D. KOLCUNOVA (2022): “Measuring Capital-Labor Substitution: The Importance of Method Choices and Publication Bias.” *Review of Economic Dynamics* **45**: pp. 55–82.
- 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.
- GINE, X., M. MARTINEZ-BRAVO, & M. VIDAL-FERNANDEZ (2017): “Are labor supply decisions consistent with neoclassical preferences? Evidence from Indian boat owners.” *Journal of Economic Behavior & Organization* **142**(C): pp. 331–347.
- GOURIO, F. & P.-A. NOUAL (2009): “The marginal worker and the aggregate elasticity of labor supply.” *Working Papers Series 2006-009*, Boston University Dept. of Economics.
- 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.
- HAM, J. C. & K. T. REILLY (2002): "Testing intertemporal substitution, implicit contracts, and hours restriction models of the labor market using micro data." *American Economic Review* **92**(4): pp. 905–927.
- 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., D. HERMAN, & Z. IRSOVA (2018a): "Does daylight saving save electricity? A meta-analysis." *The Energy Journal* **39**(2): pp. 35–62.
- HAVRANEK, T. & Z. IRSOVA (2017): "Do borders really slash trade? A meta-analysis." *IMF Economic Review* **65**(2): pp. 365–396.
- HAVRANEK, T., Z. IRSOVA, L. LASLOPOVA, & O. ZEYNALOVA (2023): "Publication and attenuation biases in measuring skill substitution." *The Review of Economics and Statistics* (forthcoming).
- HAVRANEK, T., Z. IRSOVA, & T. VLACH (2018b): "Measuring the income elasticity of water demand: the importance of publication and endogeneity biases." *Land Economics* **94**(2): pp. 259–283.
- HAVRANEK, T., M. RUSNAK, & A. SOKOLOVA (2017): "Habit formation in consumption: A meta-analysis." *European Economic Review* **95**(C): pp. 142–167.
- 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." *The Economic Journal* **131**(636): pp. 1788–1814.
- 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.
- IRSOVA, Z., P. R. D. BOM, T. HAVRANEK, & H. RACHINGER (2023): "Spurious Precision in Meta-Analysis." *CEPR Discussion Papers 17927*, Centre for Economic Policy Research.
- KARABARBOUNIS, M. (2016): "A road map for efficiently taxing heterogeneous agents." *American Economic Journal: Macroeconomics* **8**(2): pp. 182–214.
- KEANE, M. & T. NEAL (2023): "Instrument strength in IV estimation and inference: A guide to theory and practice." *Journal of Econometrics* **235**(2): pp. 1625–1653.
- 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.
- LEE, C.-I. (2001): "Finite sample bias in IV estimation."

- tion of intertemporal labor supply models: Is the intertemporal substitution elasticity really small?" *The Review of Economics and Statistics* **83**(4): pp. 638–646.
- 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* **24**(4): 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 2011-24*, Center for Retirement Research, Boston College.
- MANOLI, D. & A. WEBER (2016): "Nonparametric evidence on the effects of financial incentives on retirement decisions." *American Economic Journal: Economic Policy* **8**(4): pp. 160–182.
- MARTINEZ, I. Z., E. SAEZ, & M. SIEGENTHALER (2021): "Intertemporal labor supply substitution? Evidence from the Swiss income tax holidays." *American Economic Review* **111**(2): pp. 506–546.
- MATOUSEK, J., T. HAVRANEK, & Z. IRSOVA (2022): "Individual discount rates: a meta-analysis of experimental evidence." *Experimental Economics* **25**(1): pp. 318–358.
- 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 vendors." *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. (2019): "The effect of child support on labor supply: An estimate of the Frisch elasticity." *Working paper*, Department of Economics, Northwestern University.
- 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.
- PISTAFERRI, L. (2003): "Anticipated and unanticipated wage changes, wage risk, and intertemporal labor supply." *Journal of Labor Economics* **21**(3): pp. 729–754.
- 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 RIO, J. (2011): "The aggregate implications of individual labor supply heterogeneity." *Working paper*, Federal Research Bank of Kansas City, Research Division.
- MUSTRE-DEL RIO, 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.
- SAEZ, E. (2003): "The effect of marginal tax rates on income: A panel study of 'bracket creep'." *Journal of Public Economics* **87**(5-6): pp. 1231–1258.
- SIGURDSSON, J. (2023a): "Labor supply responses and adjustment frictions: A tax-free year in Iceland." *American Economic Journal: Economic Policy* (forthcoming).
- SIGURDSSON, J. (2023b): "The Norwegian Tax Holiday Salience, Labor Supply Responses, and Frictions." *Working paper*, Stockholm University.
- 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.
- THELOUDIS, A. (2021): “Consumption inequality across heterogeneous families.” *European Economic Review* **136(C)**: p. 103765.
- WALLENIUS, J. (2011): “Human capital accumulation and the intertemporal elasticity of substitution of labor: How large is the bias?” *Review of Economic Dynamics* **14(4)**: pp. 577–591.
- WHALEN, C. & F. REICHLING (2017): “Estimates of the Frisch elasticity of labor supply: A review.” *Eastern Economic Journal* **43(1)**: pp. 37–42.
- YANG, F., T. HAVRANEK, Z. IRSOVA, & J. NOVAK (2023): “Is Research on Hedge Fund Performance Published Selectively? A Quantitative Survey.” *Journal of Economic Surveys* (**forthcoming**).
- 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.
- ZIGRAIOVA, D., T. HAVRANEK, Z. IRSOVA, & J. NOVAK (2021): “How puzzling is the forward premium puzzle? A meta-analysis.” *European Economic Review* **134(C)**: p. 103714.
- ZILIAK, J. P. & T. J. KNIESNER (2005): “The effect of income taxation on consumption and labor supply.” *Journal of Labor Economics* **23(4)**: pp. 769–796.

Appendices (for online publication)

A Intensive Margin Elasticities

This appendix summarizes the meta-analysis of intensive margin elasticities. Our approach here is analogous to the meta-analysis of extensive margin elasticities presented in the main body of the paper, so we only briefly describe the results. All the concepts and techniques are explained in detail in the main body of the paper, especially in Section 3 and Section 4; the reader should inspect these sections before turning to this appendix. Again we use Google Scholar to search for the estimates of Frisch elasticities at the intensive margin, and the details of the search strategy are described in Figure B2. We find 40 studies, listed in Table A1, which together provide 709 estimates of the intensive margin elasticity; details on the extraction of estimates from individual studies are available in Table B1. For comparison, on the extensive margin elasticity we found 38 studies with 762 estimates, so the size of the dataset is almost the same. But for the intensive margin we only have 8 quasi-experimental studies, compared to 14 for the extensive margin. The relative scarcity of quasi-experimental evidence for the intensive margin elasticity compared to the extensive margin elasticity was noted by Chetty *et al.* (2013) and persists to this day.

As shown in Figure A1, the reported intensive margin elasticities are most commonly between 0 and 0.7, and their density is relatively flat in this interval. The mean is about 0.5 and the median 0.4. Estimated elasticities below -0.1 and above 1 are quite rare in the literature. Note the jump in the distribution at 0, which is consistent with bias against negative estimates of the elasticity; we observed a similar pattern for the extensive margin. Figure A2 shows some stylized facts in the data. Similarly to the extensive margin, estimates corresponding to workers near retirement are larger than estimates corresponding to prime-age workers. Estimates are larger for women than men and for macro data than micro data. In contrast to the extensive margin, however, for the intensive margin quasi-experimental estimates tend to be substantially larger than the rest of the micro estimates. For the intensive margin, quasi-experimental evidence does not contradict macro evidence, which was also noted by Chetty *et al.* (2013). We confirm that this finding holds with more recent data, and additionally the mean of quasi-experimental estimates (0.6) is similar to that reported by Chetty *et al.* (2013, 0.54).

But the mean of reported estimates is a misleading statistic affected in many fields (including the extensive margin Frisch elasticity, as we showed in the main body of the paper) by publication selection bias. Once again we find evidence of this bias, as apparent from Figure A3 and Table A2. The funnel plot is clearly asymmetrical, though perhaps less so than in the case of the extensive margin. All statistical tests find evidence of publication bias, and the mean elasticities corrected for this bias range between 0.2 and 0.4, with a median of 0.3. This finding implies a slightly weaker publication bias for the intensive margin compared to elasticities at the extensive margin: for both margins, the mean reported (uncorrected) elasticity is around 0.5. After correction for the bias (and ignoring for a while methodology and demographics considerations that also affect the estimates), the mean estimate is a bit smaller for the extensive margin (about 0.25) than for the intensive margin (about 0.3). One potential explanation is that with a larger underlying effect (intensive margin elasticity), less p-hacking is needed to produce statistically significant estimates.

In Table A3, we repeat the analysis of publication bias previously reported in Table A2 for two subsamples: quasi-experimental estimates and IV estimates with first-stage robust F-statistics above 10. Many authors would consider those two groups of studies as especially relevant for a proper identification of the underlying intensive margin elasticity. In addition, Keane & Neal (2023) show that for instrumental variables, estimates and standard errors are correlated by construction when instruments are weak. So we need to check whether the correlation persists even for strong instruments. (They recommend a much larger cut-off for first-stage F-statistic than the commonly used 10, but that would leave only a handful of papers in the subsample.) Even with a much reduced sample, almost all specifications in Table A2 find evidence of publication selection bias. For quasi-experimental estimates, the corrected mean effect ranges between 0 and 0.25, with a median of 0.1. For IV estimates with relatively strong instruments (first-stage F-statistics above 10), the corrected mean ranges between 0.2 and 0.6, with a median of 0.3. We conclude that evidence for publication bias is solid in the case of intensive margin elasticities, and values between 0.1 and 0.3 can be quite easily defended for the calibration of representative agent models.

Next, we focus on heterogeneity in the estimated elasticities. Table A4 summarizes the variables that reflect the context in which intensive margin elasticities are estimated; the variables

are the same as in the case of the extensive margin with the exception of a few that had to be omitted (*Ratio*, *Indivisible*, *Probit*) due to their limited variation in the intensive elasticity dataset, lack of relevance, or high correlation with other variables. The relatively modest correlations of the remaining variables are shown in Figure A4. Table A5 and Figure A5 report the results of Bayesian model averaging. BMA corroborates publication bias among intensive margin elasticities. Similarly to the extensive margin, for the intensive margin macro estimates tend to be larger than micro estimates, prime-age workers display smaller elasticities than workers near retirement, and women display larger elasticities than men. In contrast to the extensive margin, for the intensive margin data frequency can be important, recent studies tend to report estimates larger than those in older studies, estimates for the US are larger than for other countries, and quasi-experimental estimates are larger than other micro estimates. The results hold across several robustness checks, Bayesian or frequentist, reported in Table A7.

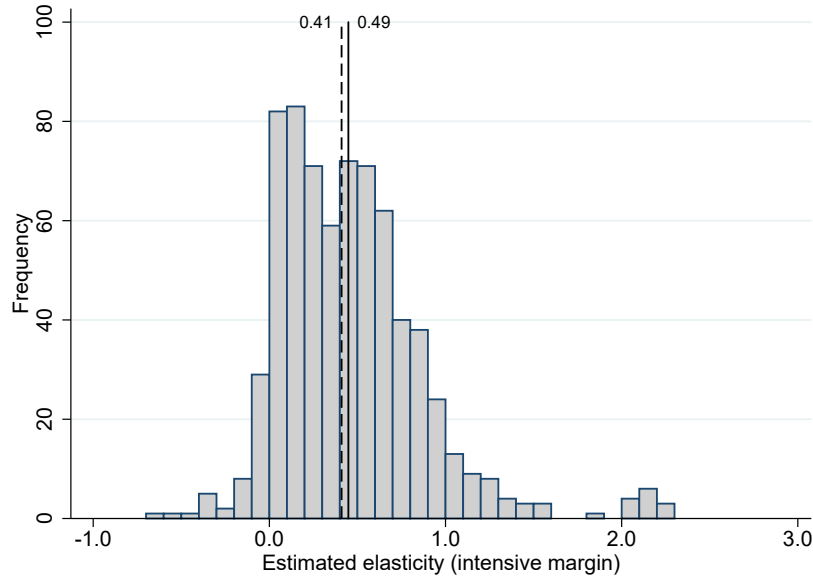
The bottom line of the meta-analysis of intensive margin elasticities is reported, together with the corresponding evidence for the extensive margin, in the main body of the paper (Table 5). The table presents implied elasticities in various contexts: that is, mean elasticities corrected for publication bias and conditional on a definition of best practice methodology. The definition is then plugged into the results of the model averaging exercise, from which fitted values for the estimated elasticities are computed. The overall mean implied elasticity at the intensive margin is 0.24 when using our subjective definition of best practice and 0.18 when defining best practice according to Martinez *et al.* (2021), a large recent quasi-experimental study published in the American Economic Review.

To avoid spurious precision, we recommend 0.2 for the calibration of the intensive margin elasticity in representative agent models. As we have noted earlier, this value is also in the middle of the interval consistent with bias-corrected means for quasi-experimental estimates and structural estimates with strong instruments. The intensive margin elasticity is larger for women and workers near retirement. Single workers seem to have smaller intensive margin elasticities, but this result should be interpreted with caution because the corresponding variable in BMA has a posterior inclusion probability smaller than 0.75, and only a small fraction of studies focus on single workers in the context of the intensive margin elasticity of intertemporal substitution in labor supply.

Table A1: Studies included in the meta-analysis of intensive margin elasticities

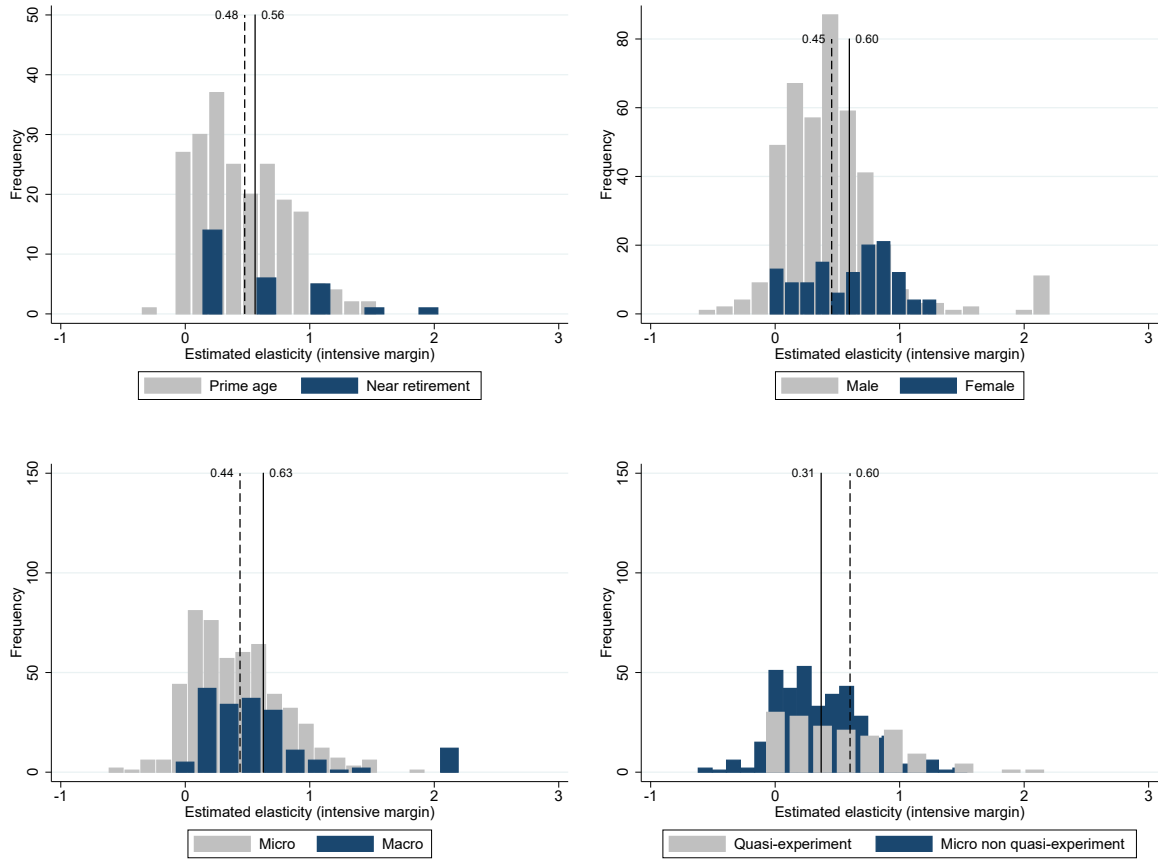
Aaronson & French (2009)	Ham & Reilly (2002)
Altonji (1986)	Inoue (2015)
Angrist (1991)	Karabarbounis (2016)
Angrist <i>et al.</i> (2021)	Keane & Wasi (2016)
Attanasio <i>et al.</i> (2018)	Kimmel & Kniesner (1998)
Battisti <i>et al.</i> (2023)	Kneip <i>et al.</i> (2019)
Beffy <i>et al.</i> (2019)	Kuroda & Yamamoto (2008)
Blundell <i>et al.</i> (2016a)	Lee (2001)
Blundell <i>et al.</i> (2016b)	Looney & Singhal (2006)
Borella <i>et al.</i> (2023)	MaCurdy (1981)
Bredemeier <i>et al.</i> (2019)	Martinez <i>et al.</i> (2021)
Caldwell & Oehlsen (2022)	Ong (2019)
Chang <i>et al.</i> (2011)	Peterman (2016)
Domeij & Floden (2006)	Pistaferri (2003)
Erosa <i>et al.</i> (2016)	Saez (2003)
Farber (2015)	Sigurdsson (2023a)
Fiorito & Zanella (2012)	Stafford (2015)
French (2005)	Theloudis (2021)
French & Stafford (2017)	Wallenius (2011)
Haan & Uhlenhorff (2013)	Ziliak & Kniesner (2005)

Figure A1: Estimates between 0 and 0.7 are almost equally common



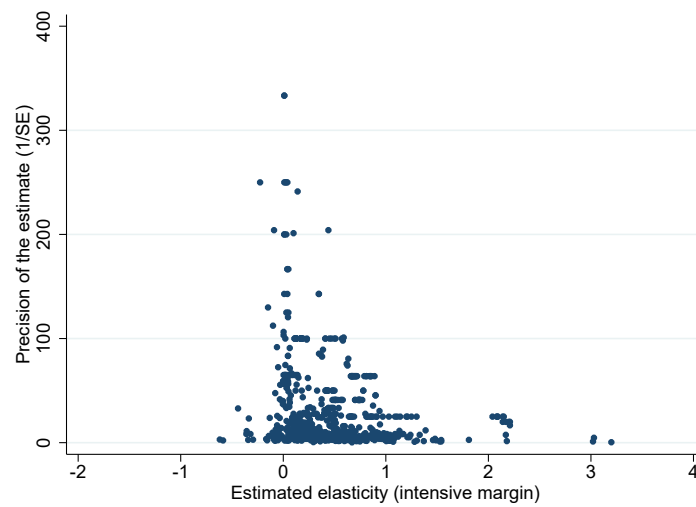
Notes: The solid line denotes the sample mean (0.49); the dashed line denotes the sample median (0.41). Note the jump at 0. Estimates smaller than -1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

Figure A2: 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.

Figure A3: The funnel plot suggests publication bias



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

Table A2: Linear and nonlinear tests document publication bias

Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias (<i>Standard error</i>)	0.590 ^{**} (0.266) [-0.01, 1.22]	0.928 ^{***} (0.110) -	1.179 ^{***} (0.440) [0.23, 2.17]	0.780 ^{***} (0.257) [0.23, 1.38]	5.163 ^{**} (2.159) {0.73, 3.72}
Effect beyond bias (<i>Constant</i>)	0.373 ^{***} (0.0567) [0.24, 0.49]	0.329 ^{***} (0.0170) -	0.297 ^{***} (0.0666) [0.14, 0.50]	0.331 ^{***} (0.0467) [0.23, 0.43]	0.279 ^{***} (0.0505) {0.04, 0.20}
First stage F-stat					9.9
Observations	709	709	709	709	663
Studies	40	40	40	40	39
Panel B: Nonlinear tests					
	Ioannidis <i>et al.</i> (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2023)
Effect beyond bias	0.199 ^{***} (0.045)	0.295 ^{***} (0.003)	0.213 ^{***} (0.014)	0.343 ^{***} (0.126)	0.387 ^{***} (0.065)
Observations	709	709	709	709	709
Studies	40	40	40	40	40

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 intensive 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. MAIVE = meta-analysis instrumental variable estimator (Irova *et al.*, 2023); 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. For MAIVE, in curly brackets we show the weak-instrument-robust Anderson-Rubin 95% confidence intervals. Panel B presents the mean elasticity corrected for publication bias using nonlinear techniques. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Publication bias in subsamples of the literature

Part 1: Quasi-experimental estimates					
Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias (<i>Standard error</i>)	1.744*** (0.630) [-0.32, 3.55]	2.036*** (0.251) -	2.803*** (0.624) [1.15, 3.92]	1.703*** (0.521) [0.63, 3.43]	8.698* (4.616) {-0.35, 17.75}
Effect beyond bias (<i>Constant</i>)	0.224** (0.111) [0.04, 0.59]	0.176*** (0.0451) -	0.0513 (0.0470) [-0.78, 0.51]	0.224*** (0.0610) [0.08, 0.52]	0.121 (0.0837) {-0.004, 0.25}
First stage F-stat					3.1
Observations	162	162	162	162	132
Studies	8	8	8	8	8
Panel B: Nonlinear tests					
	Ioannidis <i>et al.</i> (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2023)
Effect beyond bias	0.028 (NA)	-0.027 (0.02)	-0.002 (0.008)	0.234* (0.122)	0.155 (0.504)
Observations	162	162	162	162	162
Studies	8	8	8	8	8
Part 2: IV estimates with first-stage F-statistics > 10					
Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias (<i>Standard error</i>)	0.523** (0.239) [-0.10, 0.93]	0.327 (0.335) -	0.728** (0.309) [-0.20, 0.96]	0.692** (0.293) [-0.13, 1.20]	-3.393* (1.821) {-8.98, -0.69}
Effect beyond bias (<i>Constant</i>)	0.285*** (0.0586) [0.21, 0.45]	0.327*** (0.0724) -	0.246*** (0.0714) [0.19, 0.48]	0.262*** (0.0620) [0.21, 0.45]	0.587*** (0.110) {0.12, 1.55}
First stage F-stat					19.2
Observations	92	92	92	92	92
Studies	6	6	6	6	6
Panel B: Nonlinear tests					
	Ioannidis <i>et al.</i> (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2023)
Effect beyond bias	0.247** (0.112)	0.421*** (0.06)	0.204*** (0.055)	0.277** (0.121)	0.375*** (0.145)
Observations	92	92	92	92	92
Studies	6	6	6	6	6

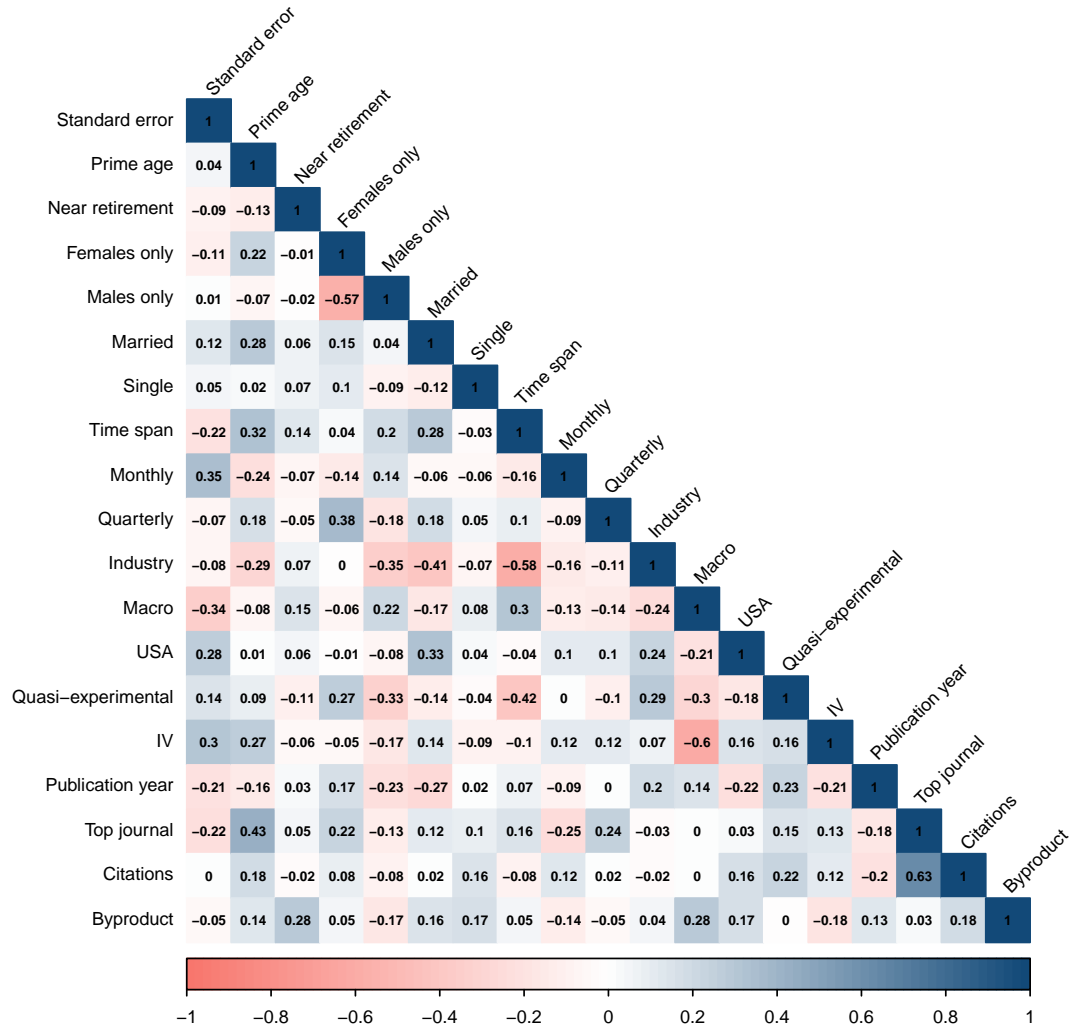
Notes: Panel A presents the results of regression $\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}$, where $\hat{\eta}_{ij}$ and $SE(\hat{\eta}_{ij})$ are the i -th estimated Frisch intensive margin elasticity and its standard error reported in the j -th study. OLS = ordinary least 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. MAIVE = meta-analysis instrumental variable estimator (Irsova *et al.*, 2023); 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). In square brackets we report 95% confidence intervals from wild bootstrap clustering. In curly brackets we show the Anderson-Rubin 95% confidence intervals. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
Frisch elasticity	Estimate of the intensive margin Frisch elasticity (response variable).	0.49	0.55
Standard error (SE)	Standard error of the estimate (the variable is important for gauging publication bias).	0.15	0.22
<i>Demographic characteristics</i>			
Prime age	= 1 if the sample only consists of people between 25 and 55 years of age.	0.30	0.46
Near retirement	= 1 if the sample only consists of people older than 55.	0.04	0.19
Females only	= 1 if the sample consists of females only.	0.18	0.38
Males only	= 1 if the sample consists of males only.	0.60	0.49
Married	= 1 if the sample consists of married people only.	0.47	0.50
Single	= 1 if the sample consists of single people only.	0.02	0.15
<i>Data characteristics</i>			
Time span	The logarithm of the data time span used to estimate the elasticity.	2.55	0.84
Monthly	= 1 if the data frequency is monthly (reference category: annual).	0.12	0.32
Quarterly	= 1 if the data frequency is quarterly (reference category: annual).	0.06	0.23
Industry	= 1 if the sample consists of workers in a specific industry (reference category: whole economy data).	0.16	0.37
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.	0.77	0.42
<i>Specification characteristics</i>			
Quasi-experimental	= 1 if the estimation framework uses quasi-experimental identification.	0.23	0.42
IV	= 1 if instrumental variable methods are used for the estimate (reference category: OLS).	0.56	0.50
<i>Publication characteristics</i>			
Publication year	The logarithm of the year the study was published.	3.42	0.53
Top journal	= 1 if the estimate is published in a top five journal in economics.	0.32	0.47
Citations	The logarithm of the number of per-year citations of the study in Google Scholar.	2.05	1.42
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.	0.13	0.33

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

Figure A4: Correlations among explanatory variables



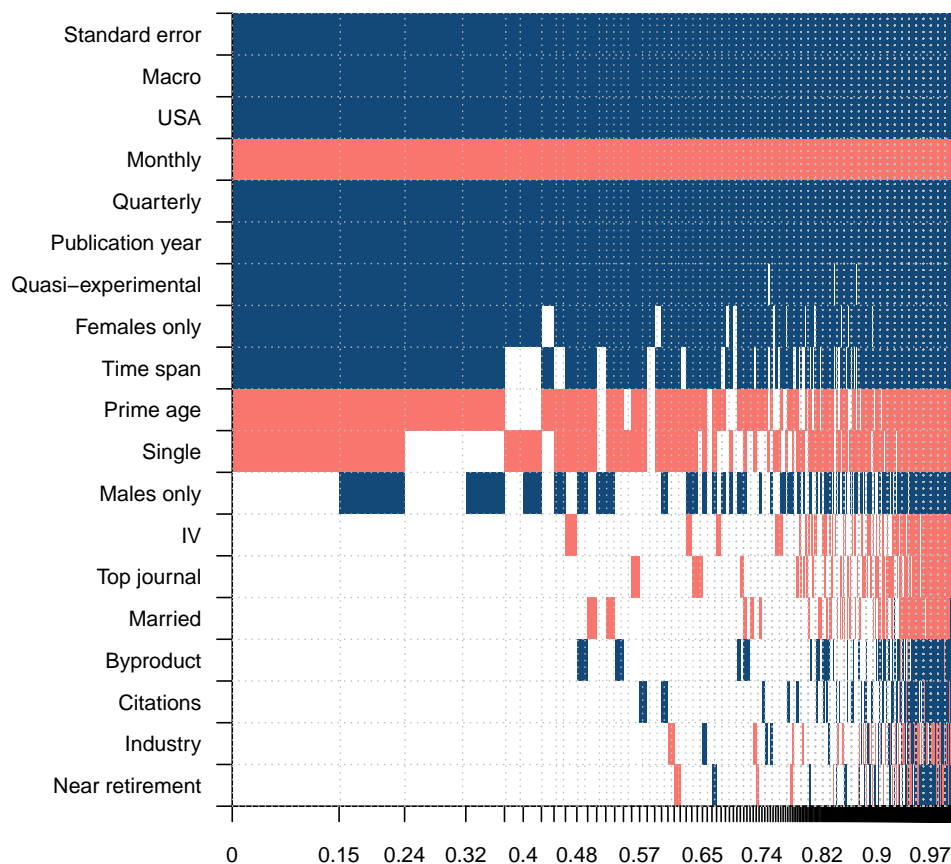
Notes: The figure shows Pearson correlation coefficients for the variables described in Table A4; only intensive margin elasticities are used for the computation.

Table A5: Why do estimates of the elasticity vary?

Response variable: Frisch elasticity (intensive margin)	Bayesian model averaging (baseline model)			Ordinary least squares (frequentist check)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Intercept	-0.405	NA	1.000	-0.391	0.190	0.046
Standard error	1.025	0.104	1.000	1.022	0.222	0.000
<i>Demographic characteristics</i>						
Prime age	-0.073	0.047	0.787	-0.098	0.062	0.122
Near retirement	0.001	0.016	0.060			
Females only	0.122	0.055	0.924	0.106	0.067	0.122
Males only	0.028	0.039	0.408			
Married	-0.002	0.012	0.089			
Single	-0.137	0.115	0.665			
<i>Data characteristics</i>						
Time span	0.044	0.028	0.799	0.062	0.038	0.112
Monthly	-0.190	0.040	1.000	-0.185	0.071	0.013
Quarterly	0.261	0.058	0.999	0.260	0.193	0.186
Industry	-0.001	0.017	0.075			
Macro	0.252	0.032	1.000	0.251	0.066	0.001
USA	0.208	0.030	1.000	0.203	0.070	0.006
<i>Specification characteristics</i>						
Quasi-experimental	0.157	0.045	0.988	0.171	0.080	0.039
IV	-0.003	0.014	0.101			
<i>Publication characteristics</i>						
Publication year	0.101	0.029	0.991	0.090	0.046	0.060
Top journal	-0.003	0.015	0.096			
Citations	0.001	0.004	0.076			
Byproduct	0.003	0.016	0.089			
Observations	709			709		
Studies	40			40		

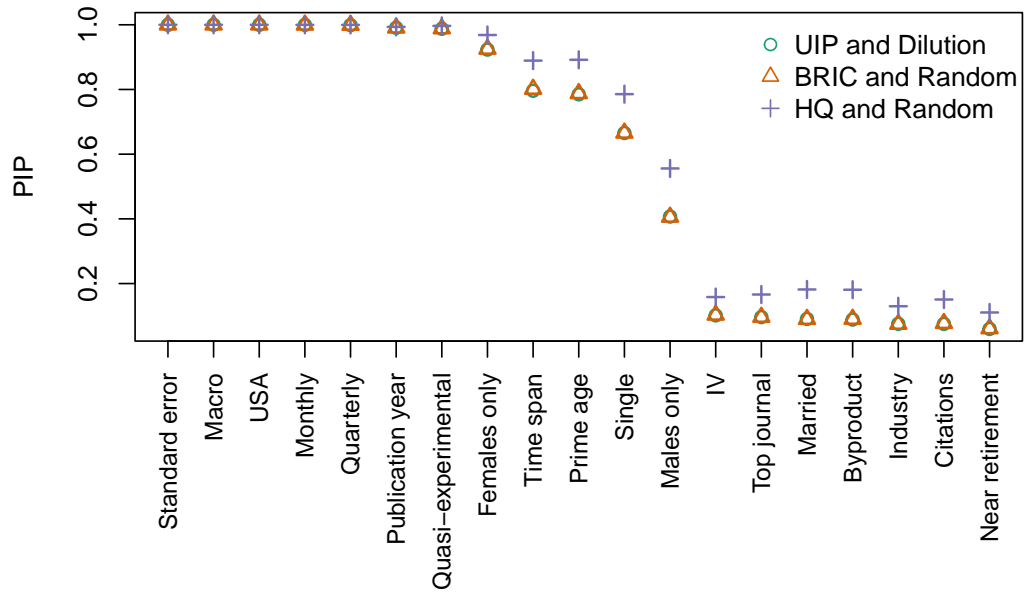
Notes: The response variable is the Frisch elasticity of labor supply at the intensive margin. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = Posterior inclusion probability, SE = standard error. The left-hand panel applies BMA based on the UIP g-prior and the dilution prior (Eicher *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 A4 presents a detailed description of all the variables.

Figure A5: Model inclusion in Bayesian model averaging (UIP and dilution prior)



Notes: The response variable is the reported estimate of the Frisch elasticity of labor supply at the intensive 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 A4 presents a detailed description of all variables. The numerical results are reported in Table A7.

Figure A6: Posterior inclusion probabilities hold across different priors



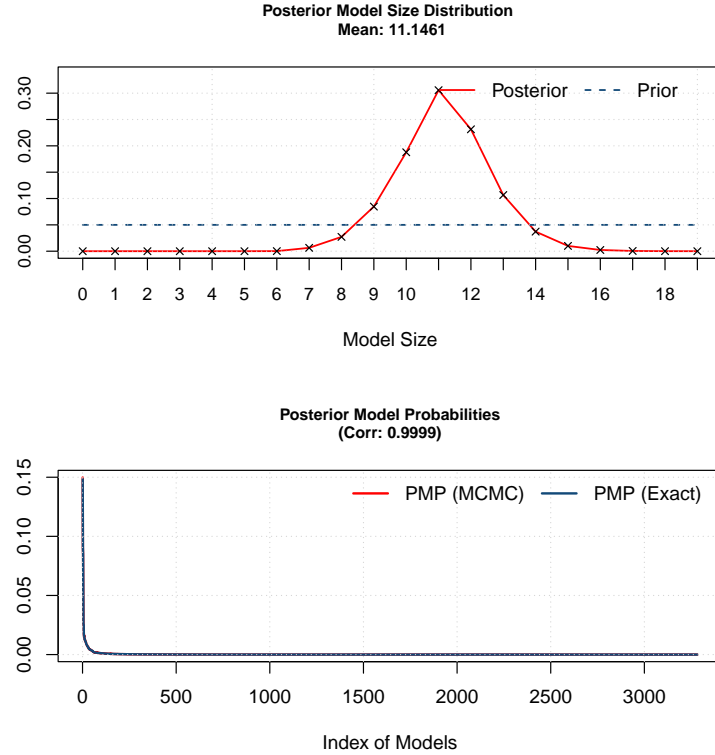
Notes: UIP and Dilution = priors according to Eicher *et al.* (2011) and George (2010). BRIC and Random = the benchmark g-prior for parameters with the beta-binomial model prior (each model size has equal prior probability). The HQ prior asymptotically mimics the Hannan-Quinn criterion. PIP = posterior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

Table A6: Summary of the BMA estimation (UIP and dilution prior)

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
11.1461	$3 \cdot 10^6$	$1 \cdot 10^6$	12.08 mins	688,859
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$5.24 \cdot 10^5$	131.0%	100%	0.9999	709
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/9.5	UIP	$A_v = 0.9986$		

Notes: The results of this BMA specification are reported in Table A5. 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 A7: Model size and convergence in the BMA model (UIP and dilution prior)



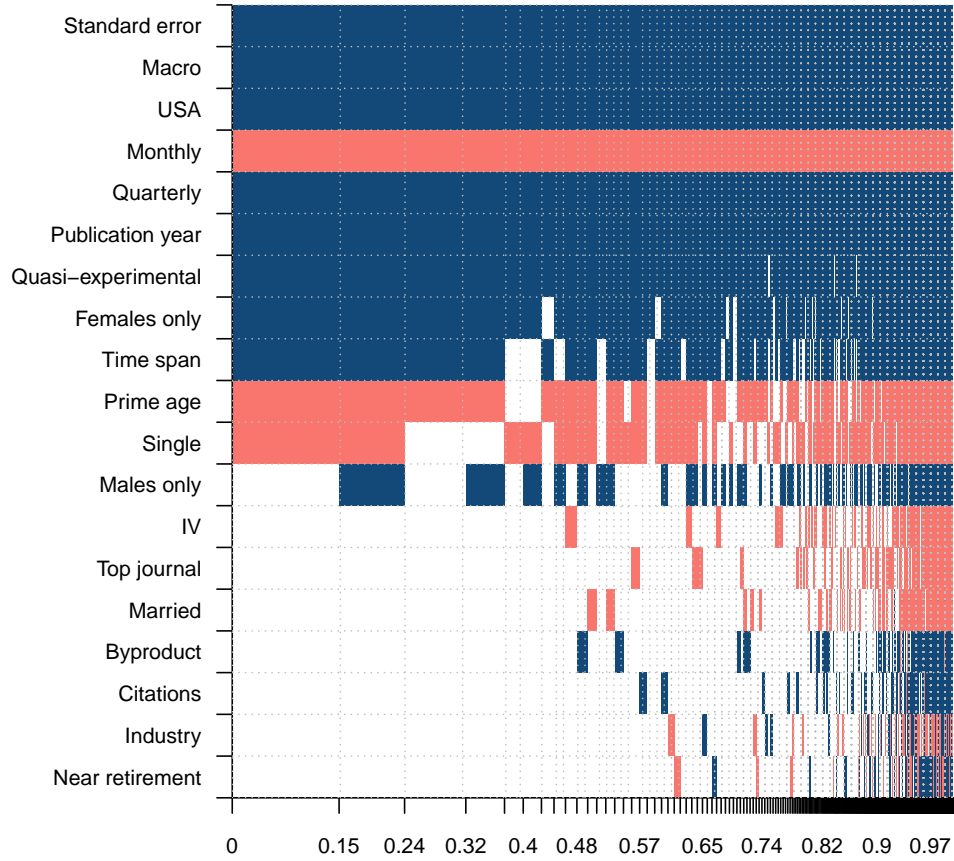
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table A5.

Table A7: Results of BMA with alternative priors and results of FMA

Response variable: Frisch elasticity (intensive margin)	Bayesian model averaging (BRIC g-prior)			Bayesian model averaging (HQ g-prior)			Frequentist model averaging		
	P. mean	P. SD	PIP	P. mean	P. SD	PIP	Coef.	SE	p-value
Intercept	-0.405	NA	1.000	-0.413	NA	1.000	-0.417	0.114	0.000
Standard error	1.025	0.104	1.000	1.032	0.104	1.000	1.042	0.111	0.000
<i>Demographic characteristics</i>									
Prime age	-0.073	0.047	0.788	-0.083	0.041	0.892	-0.089	0.034	0.008
Near retirement	0.001	0.016	0.061	0.001	0.021	0.111	0.010	0.064	0.876
Females only	0.122	0.055	0.925	0.135	0.050	0.968	0.169	0.043	0.000
Males only	0.028	0.039	0.406	0.038	0.041	0.556	0.071	0.032	0.029
Married	-0.002	0.012	0.090	-0.006	0.018	0.182	-0.043	0.033	0.188
Single	-0.137	0.115	0.666	-0.162	0.108	0.786	-0.238	0.082	0.004
<i>Data characteristics</i>									
Time span	0.045	0.028	0.801	0.049	0.025	0.889	0.061	0.022	0.005
Monthly	-0.190	0.040	1.000	-0.195	0.040	1.000	-0.221	0.044	0.000
Quarterly	0.261	0.058	0.999	0.261	0.057	1.000	0.279	0.058	0.000
Industry	-0.001	0.017	0.075	-0.001	0.022	0.130	-0.009	0.060	0.876
Macro	0.252	0.032	1.000	0.246	0.033	1.000	0.211	0.039	0.000
USA	0.208	0.030	1.000	0.208	0.031	1.000	0.211	0.040	0.000
<i>Specification characteristics</i>									
Quasi-experimental	0.157	0.045	0.988	0.161	0.041	0.997	0.164	0.038	0.000
IV	-0.003	0.015	0.103	-0.004	0.016	0.158	-0.018	0.031	0.552
<i>Publication characteristics</i>									
Publication year	0.101	0.029	0.991	0.098	0.028	0.994	0.089	0.028	0.001
Top journal	-0.003	0.015	0.096	-0.006	0.020	0.166	-0.044	0.041	0.280
Citations	0.001	0.004	0.077	0.001	0.006	0.151	0.014	0.013	0.273
Byproduct	0.003	0.016	0.090	0.008	0.024	0.181	0.046	0.044	0.287
Observations	709			709			709		
Studies	40			40			40		

Notes: The response variable is the Frisch elasticity of labor supply at the intensive margin. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = Posterior inclusion probability, SE = standard error. In the left-hand panel we apply BMA based on BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior). The middle panel reports the results of BMA based on HQ g-prior, which asymptotically mimics the Hannan-Quinn criterion. Table A4 presents a detailed description of all variables. In the right-hand panel we use Mallow's weights Hansen (2007) and the orthogonalization of the covariate space suggested by Amini & Parmeter (2012) to conduct the frequentist model averaging exercise.

Figure A8: Model inclusion in Bayesian model averaging (Random and BRIC)



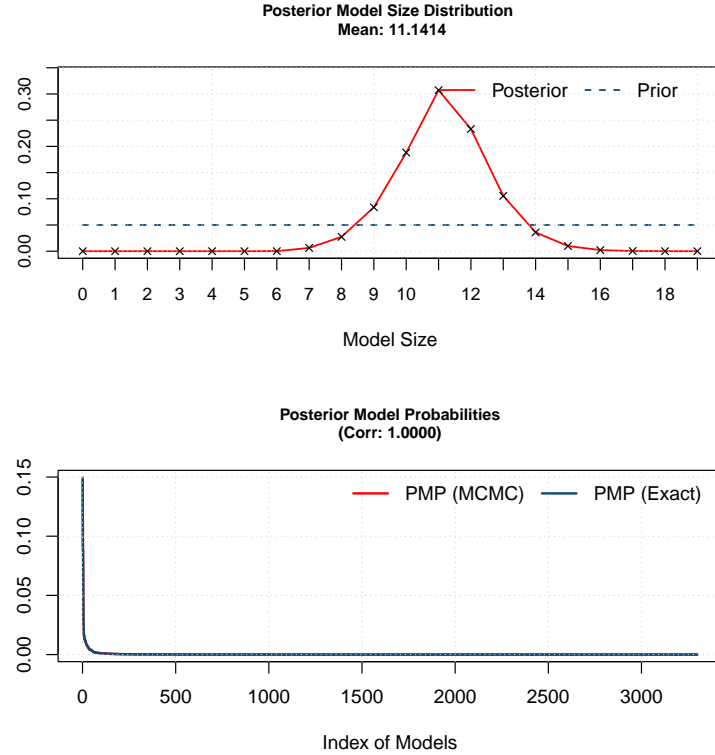
Notes: The response variable is the estimate of the Frisch elasticity of labor supply at the intensive margin. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The estimation is based on 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 A7.

Table A8: Summary of the BMA (Random and BRIC)

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
11.1414	$3 \cdot 10^6$	$1 \cdot 10^6$	12.05 mins	684,908
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$5.24 \cdot 10^5$	131.0%	100%	1.0000	709
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/9.5	BRIC	$A_v = 0.9986$		

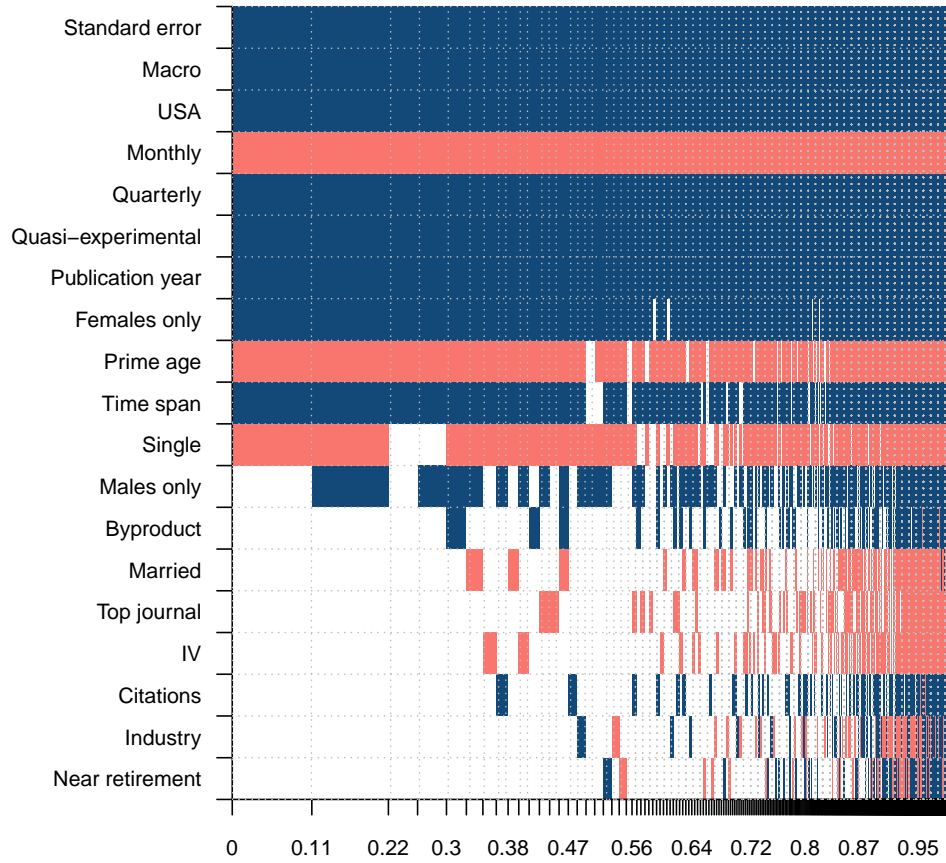
Notes: The results of this BMA specification are reported in Table A7. The estimation is based on BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure A9: Model size and convergence in the BMA (Random and BRIC)



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA (random and BRIC prior) exercise reported in Table A7.

Figure A10: Model inclusion in BMA (Random and HQ g-prior)



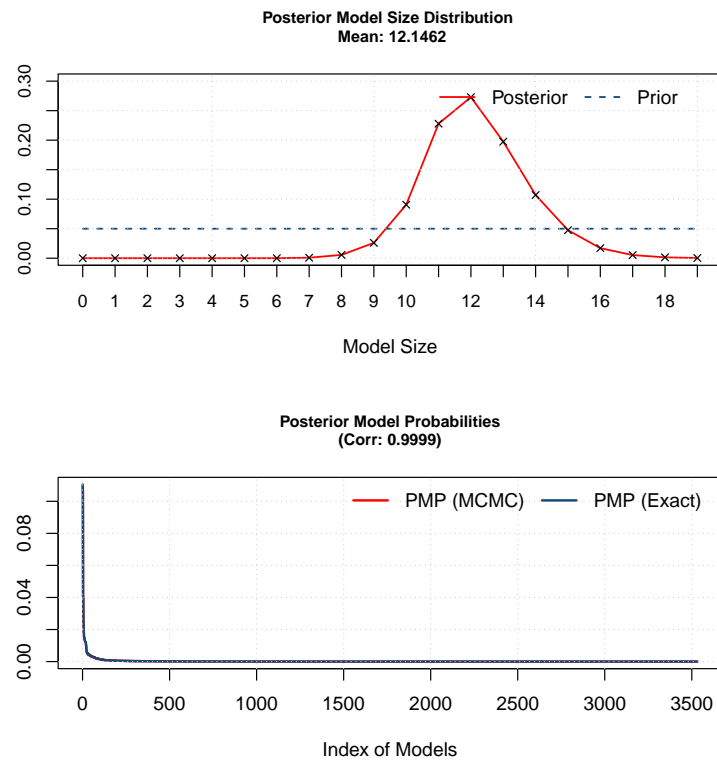
Notes: The response variable is the estimate of the Frisch intensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 A7.

Table A9: Summary of the BMA (Random and HQ g-prior)

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
12.1462	$3 \cdot 10^6$	$1 \cdot 10^6$	13.61 mins	801,966
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$5.24 \cdot 10^5$	153.0%	100%	1.0000	709
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/9.5	Hannan-Quinn	Av = 0.9965		

Notes: The results of this BMA specification are reported in Table A7. The estimation is based on HQ g-prior that asymptotically mimics the Hannan-Quinn criterion and random model prior as suggested by Fernandez *et al.* (2001).

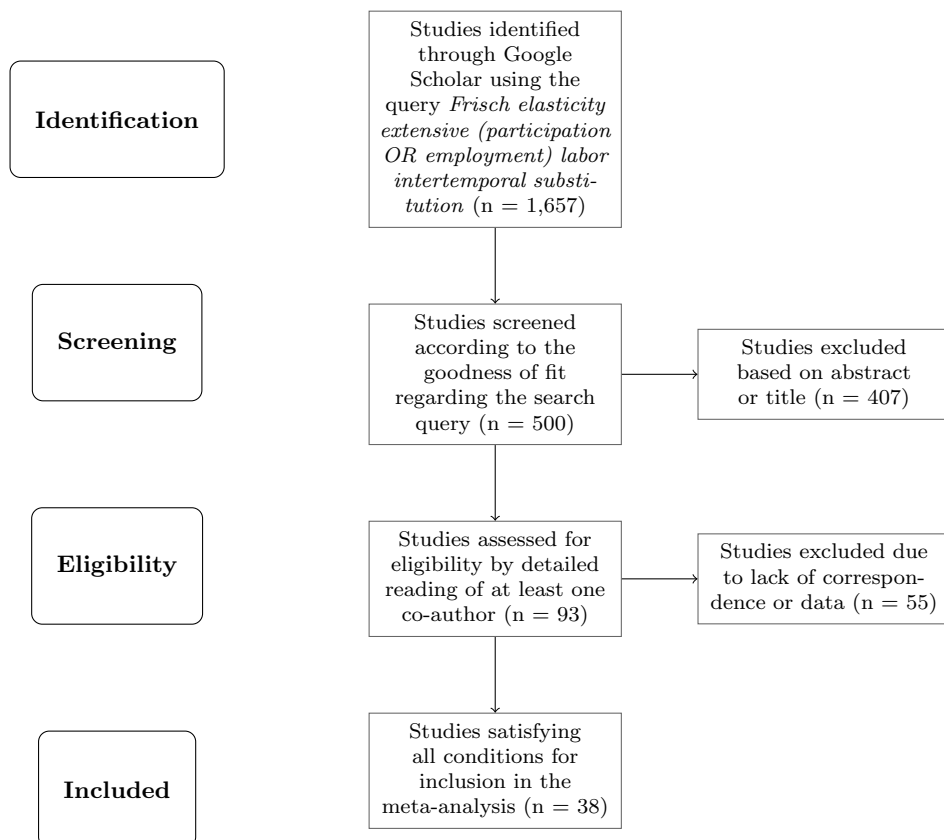
Figure A11: Model size and convergence in the BMA (Random and HQ g-prior)



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA (random and HQ g-prior) exercise reported in Table A7.

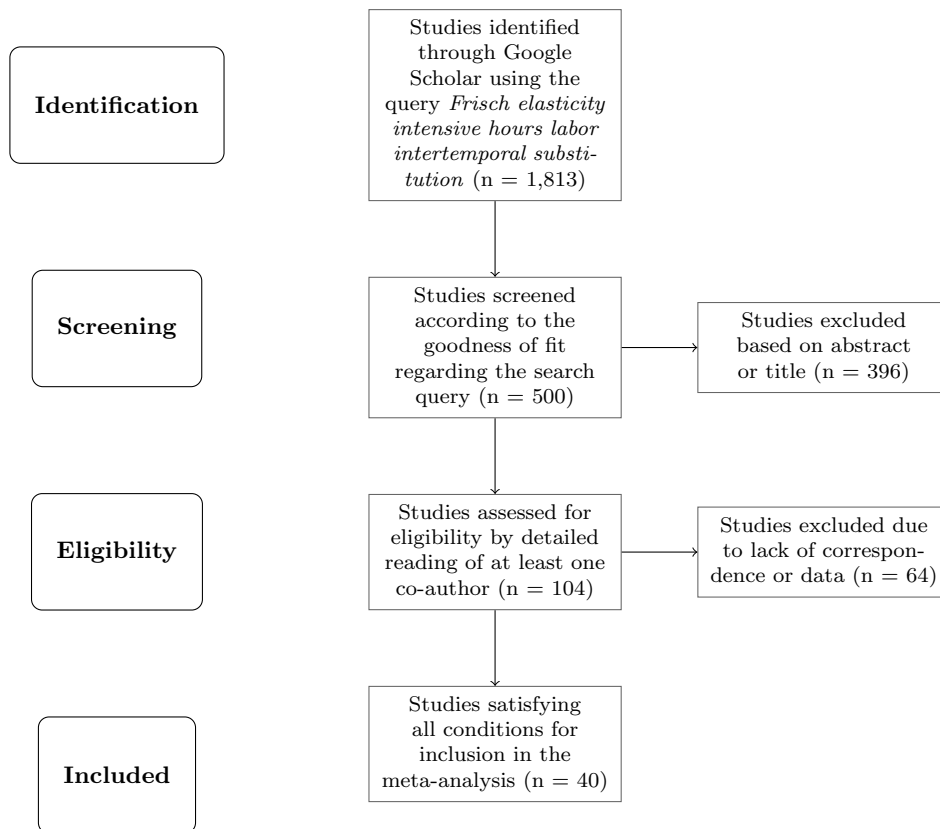
B Details on Literature Search and Data Collection

Figure B1: The PRISMA flow diagram (extensive margin elasticities)



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).

Figure B2: The PRISMA flow diagram (intensive margin elasticities)



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).

Table B1: Sources for estimates collected from individual papers

Extensive margin	Source	Intensive margin	Source
Attanasio <i>et al.</i> (2018)	Tables VIII-X ¹	Aaronson & French (2009)	Tables 2-3
Beffy <i>et al.</i> (2019)	Table 11	Altonji (1986)	Tables 1-2, 4
Bianchi <i>et al.</i> (2001)	Tables 4-6, 8	Angrist (1991)	Tables 2, 4
Blundell <i>et al.</i> (2016a)	Table XIV	Angrist <i>et al.</i> (2021)	Table 5
Blundell <i>et al.</i> (2016b)	Table 7	Attanasio <i>et al.</i> (2018)	Table VIII-X
Borella <i>et al.</i> (2023)	Table 4	Battisti <i>et al.</i> (2023)	Table 5
Brown (2013)	Via Chetty <i>et al.</i> (2013) ²	Beffy <i>et al.</i> (2019)	Table 11
Caldwell (2019)	Table 3.7	Blundell <i>et al.</i> (2016a)	Table XIV
Card & Hyslop (2005)	Via Chetty <i>et al.</i> (2013) ³	Blundell <i>et al.</i> (2016b)	Tables 4-6
Carrington (1996)	Table 2	Borella <i>et al.</i> (2023)	Table 4
Chang & Kim (2006)	Table 8	Bredemeier <i>et al.</i> (2019)	Tables 1-5, B2-F4
Chang <i>et al.</i> (2019)	Table 7	Caldwell & Oehlsen (2022)	Tables 4, A6-7
Erosa <i>et al.</i> (2016)	Tables 4-5	Chang <i>et al.</i> (2011)	Table 1
Espino <i>et al.</i> (2017)	Table 4	Domeij & Floden (2006)	Tables 2, 4-7
Fiorito & Zanella (2012)	Table 3, 6 ⁴	Erosa <i>et al.</i> (2016)	Table 4
French & Stafford (2017)	Tables 2-3	Farber (2015)	Tables IV-VI
Gine <i>et al.</i> (2017)	Table 6	Fiorito & Zanella (2012)	Table 6
Gourio & Noual (2009)	Abstract and Table 7	French (2005)	Tables 2, 5
Gruber & Wise (1999)	Via Chetty <i>et al.</i> (2013) ⁵	French & Stafford (2017)	Tables 2-3
Haan & Uhlenhorff (2013)	Table 6	Haan & Uhlenhorff (2013)	Table 6
Inoue (2015)	Tables 3-6	Ham & Reilly (2002)	Table 1
Karabarbounis (2016)	Table 3	Inoue (2015)	Tables 3-6
Keane & Wasi (2016)	Figure 19 ⁶	Karabarbounis (2016)	Table 3
Kimmel & Kniesner (1998)	Table 1	Keane & Wasi (2016)	Figure 20 ⁷
Kneip <i>et al.</i> (2019)	Tables 3, E.2, F.1-3	Kimmel & Kniesner (1998)	Table 1
Kuroda & Yamamoto (2008)	Tables 2-5 ⁸	Kneip <i>et al.</i> (2019)	Tables 3, D.2, E.2, F.1-3
Looney & Singhal (2006)	Table 36	Kuroda & Yamamoto (2008)	Tables 3, 5
Manoli & Weber (2011)	Tables 3-4, 5A-B	Lee (2001)	Tables 1-2
Manoli & Weber (2016)	Table 3	Looney & Singhal (2006)	Tables 5, 8
Martinez <i>et al.</i> (2021)	Tables 3-4	MaCurdy (1981)	Table 1
Mustre-del Rio (2011)	Table 5	Martinez <i>et al.</i> (2021)	Tables 2-5
Mustre-del Rio (2015)	Table 8	Ong (2019)	Tables 2, A2
Oettinger (1999)	Table 5	Peterman (2016)	Tables 2-4, 9
Ong (2019)	Tables 2-3, A3	Pistaferri (2003)	Tables 2-3
Park (2020)	Tables 1, 8	Saez (2003)	Tables 5-6
Peterman (2016)	Table 5	Sigurdsson (2023a)	Tables 1, A.1
Sigurdsson (2023a)	Tables 2, A.10, A.28	Stafford (2015)	Tables 2, 4
Stafford (2015)	Tables 2, 4	Theloudis (2021)	Table 4
		Wallenius (2011)	Tables 1-3
		Ziliak & Kniesner (2005)	Tables 2-3

¹The difference between reported aggregate elasticities and median intensive elasticities.²Computed based on the approach described in Chetty *et al.* (2013).³Computed based on the approach described in Chetty *et al.* (2013).⁴The difference between aggregate elasticities in Table 3 and pure intensive elasticities in Table 6.⁵Computed based on the approach described in Chetty *et al.* (2013).⁶Elasticity of employment for ages 25, 40, and 55 with a college education.⁷Elasticity of employment for ages 25, 40, and 55 with a college education.⁸The difference between aggregate and intensive elasticities in Tables 2-3 and 4-5.

C Estimating the Elasticities

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 4. 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:

$$\eta = \frac{\partial h_t}{\partial w_t} \frac{w_t}{h_t} \Big|_{\lambda}, \quad (2)$$

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:

$$\ln h_t = \alpha_i + \rho + \theta x_t + \eta \ln w_t + \varepsilon_t, \quad (3)$$

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

$$h_t = \begin{cases} \bar{h}, & \text{if } w_t \geq w_t^R \\ 0, & \text{if } w_t \leq w_t^R. \end{cases} \quad (4)$$

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:

$$\text{var}(\ln H_t) = \text{var}(\ln n_t) + \text{var}(\ln \bar{h}_t) + 2 \text{cov}(\ln n_t, \ln \bar{h}_t), \quad (5)$$

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 (5), the decomposition of aggregate Frisch elasticity can be written as

$$\eta = \frac{\text{cov}(\Delta \ln H, \Delta \ln W)}{\text{var}(\Delta \ln W)} = \frac{\text{cov}(\Delta \ln \bar{h}, \Delta \ln W)}{\text{var}(\Delta \ln W)} + \frac{\text{cov}(\Delta \ln n, \Delta \ln W)}{\text{var}(\Delta \ln W)}, \quad (6)$$

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 4.

D Diagnostics and Robustness Checks of the Meta-Analysis of Extensive Margin Elasticities

Table D1: Publication bias tests in a subsample of quasi-experimental estimates

Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias (<i>Standard error</i>)	0.992** (0.488) [-0.20, 2.92]	0.0415 (0.283) -	1.479** (0.720) [-3.12, 7.74]	1.498** (0.683) [0.23, 3.13]	0.643 (0.460) {-0.04, 2.33}
Effect beyond bias (<i>Constant</i>)	0.153*** (0.0469) [-0.01, 0.28]	0.211*** (0.0213) -	0.123*** (0.0467) [-0.01, 0.22]	0.170*** (0.0479) [0.05, 0.29]	0.188*** (0.0393) {-0.01, 0.68}
First stage F-stat					10.3
Observations	202	202	202	202	179
Studies	14	14	14	14	13
Panel B: Nonlinear tests					
	Ioannidis <i>et al.</i> (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2023)
Effect beyond bias	0.112** (0.049)	0.211*** (0.048)	0.083*** (0.015)	0.095 (0.082)	0.217*** (0.057)
Observations	202	202	202	202	202
Studies	14	14	14	14	14

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. MAIVE = meta-analysis instrumental variable estimator (Irsova *et al.*, 2023); 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. In curly brackets we show the Anderson-Rubin 95% confidence interval. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D2: Correlation between elasticities and standard errors is weaker for stronger instruments

	OLS
Standard error (SE)	1.876*** (0.518)
SE * First-stage F-stat	-0.0110** (0.00430)
Constant	0.133* (0.0725)
Observations	22
Studies	4

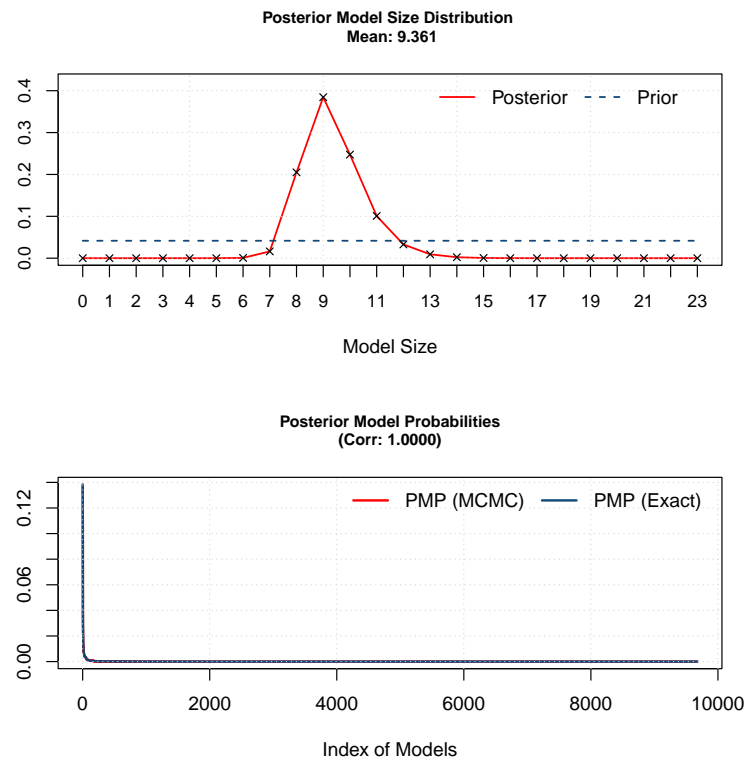
Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D3: Summary of the benchmark BMA estimation

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
9.361	$3 \cdot 10^6$	$1 \cdot 10^6$	12.89 mins	546,667
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$8.39 \cdot 10^6$	6.5%	100%	1.0000	762
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/11.5	UIP	$A_v = 0.9987$		

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 D1: Model size and convergence in the benchmark BMA model



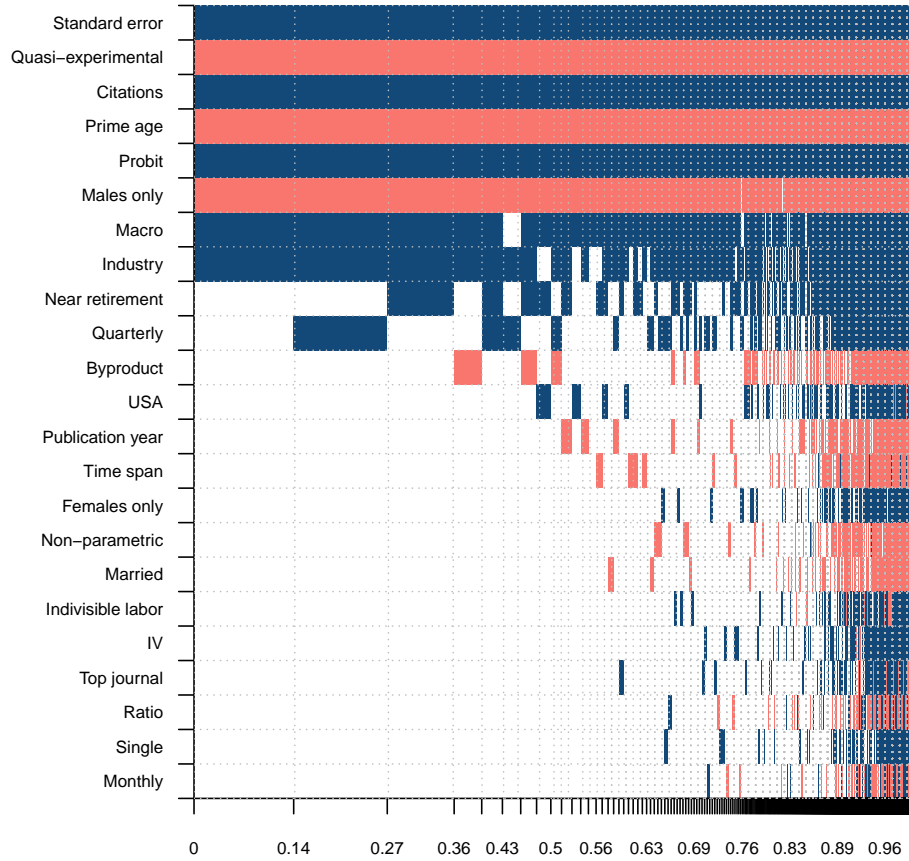
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table 4.

Table D4: Results of BMA with alternative priors and results of FMA

Response variable: Frisch elasticity (extensive margin)	Bayesian model averaging (BRIC g-prior)			Bayesian model averaging (HQ g-prior)			Frequentist model averaging		
	P. mean	P. SD	PIP	P. mean	P. SD	PIP	Coef.	SE	p-value
Intercept	0.326	NA	1.000	0.353	NA	1.000	0.876	0.310	0.005
Standard error	1.381	0.120	1.000	1.371	0.124	1.000	1.254	0.173	0.000
<i>Demographic characteristics</i>									
Prime age	-0.150	0.030	1.000	-0.146	0.031	1.000	-0.127	0.033	0.000
Near retirement	0.034	0.047	0.389	0.047	0.051	0.535	0.112	0.038	0.003
Females only	0.003	0.014	0.057	0.005	0.020	0.109	0.089	0.038	0.017
Males only	-0.113	0.032	0.980	-0.113	0.033	0.976	-0.057	0.038	0.130
Married	-0.002	0.015	0.047	-0.004	0.018	0.079	-0.019	0.048	0.697
Single	0.001	0.012	0.035	0.003	0.017	0.068	0.072	0.054	0.183
<i>Data characteristics</i>									
Time span	-0.002	0.010	0.074	-0.002	0.010	0.098	0.032	0.028	0.239
Monthly	0.000	0.015	0.029	0.000	0.020	0.054	0.004	0.083	0.963
Quarterly	0.030	0.045	0.363	0.032	0.044	0.411	0.103	0.048	0.030
Ratio	0.000	0.008	0.037	0.000	0.010	0.063	0.052	0.041	0.200
Industry	0.129	0.066	0.859	0.134	0.064	0.886	0.297	0.088	0.001
Macro	0.134	0.051	0.942	0.140	0.049	0.964	0.217	0.051	0.000
USA	0.007	0.023	0.111	0.007	0.024	0.137	-0.014	0.044	0.757
<i>Specification characteristics</i>									
Indivisible labor	0.002	0.013	0.045	0.004	0.021	0.088	0.109	0.058	0.062
Quasi-experimental	-0.285	0.042	1.000	-0.287	0.042	1.000	-0.277	0.058	0.000
Probit	0.232	0.057	0.995	0.229	0.057	0.996	0.178	0.065	0.006
Non-parametric	-0.002	0.014	0.056	-0.006	0.022	0.118	-0.062	0.052	0.239
IV	0.001	0.012	0.042	0.003	0.017	0.080	0.034	0.057	0.559
<i>Publication characteristics</i>									
Publication year	-0.010	0.039	0.089	-0.018	0.052	0.158	-0.232	0.098	0.018
Top journal	0.001	0.010	0.040	0.002	0.013	0.071	-0.014	0.045	0.754
Citations	0.067	0.013	1.000	0.067	0.013	1.000	0.070	0.016	0.000
Byproduct	-0.016	0.042	0.164	-0.026	0.051	0.266	-0.127	0.055	0.022
Observations	762			762			762		
Studies	38			38			38		

Notes: The response variable is the Frisch elasticity of labor supply at the extensive margin. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = Posterior inclusion probability, SE = standard error. In the left-hand panel we apply BMA based on BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior). The middle 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. In the right-hand panel we use Mallows weights Hansen (2007) and the orthogonalization of the covariate space suggested by Amini & Parmeter (2012) to conduct the frequentist model averaging exercise.

Figure D2: Model inclusion in BMA (BRIC g-prior)



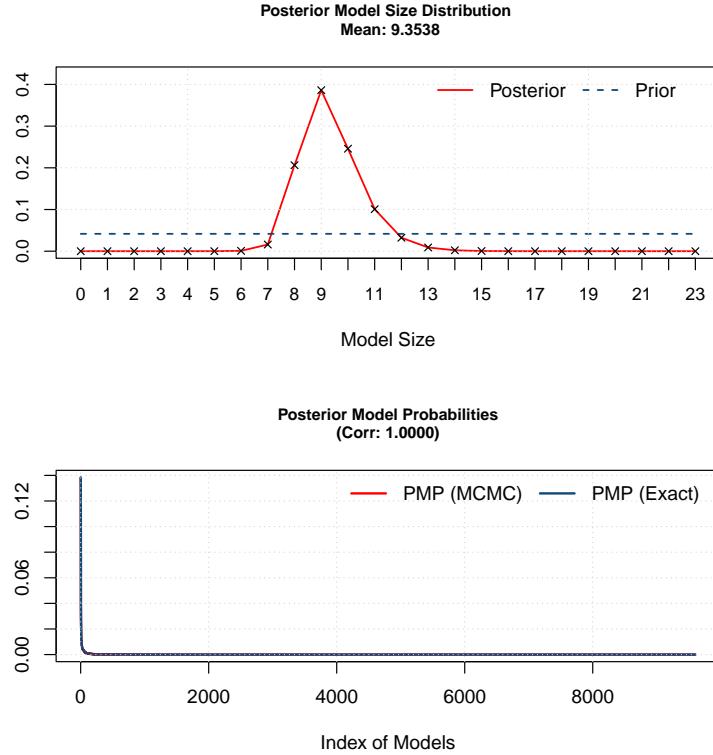
Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 D4.

Table D5: Summary of the BMA (BRIC g-prior)

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
9.3538	$3 \cdot 10^6$	$1 \cdot 10^6$	13.07 mins	544,779
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$8.39 \cdot 10^6$	6.5%	100%	1.0000	762
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/11.5	BRIC	$A_v = 0.9987$		

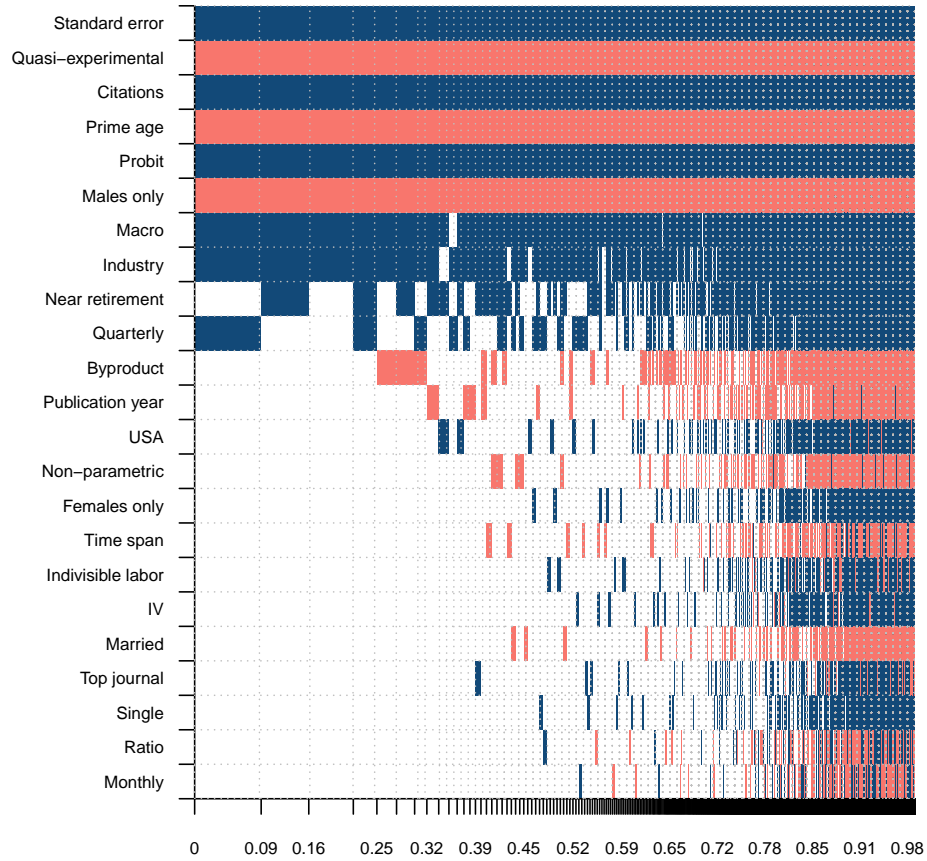
Notes: The results of this BMA specification are reported in Table D4. The estimation is based on BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure D3: Model size and convergence in the BMA (BRIC g-prior)



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA (random and BRIC prior) exercise reported in Table D4.

Figure D4: Model inclusion in BMA (Random and HQ g-prior)



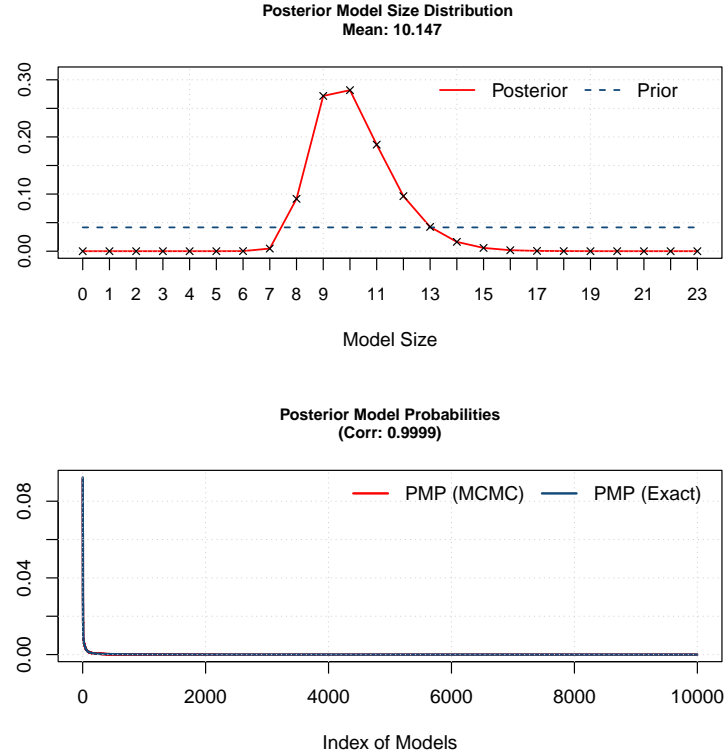
Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 D4.

Table D6: Summary of the BMA (Random and HQ g-prior)

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
10.147	$3 \cdot 10^6$	$1 \cdot 10^6$	16.38 mins	718,854
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$8.39 \cdot 10^6$	8.6%	99%	0.9999	762
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Random/11.5	Hannan-Quinn	Av = 0.9966		

Notes: The results of this BMA specification are reported in Table D4. The estimation is based on HQ g-prior that asymptotically mimics the Hannan-Quinn criterion and random model prior as suggested by Fernandez *et al.* (2001).

Figure D5: Model size and convergence in the BMA (Random and HQ g-prior)



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA (random and HQ g-prior) exercise reported in Table D4.