CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Essays on the Meta-Analysis of Deep Parameters

Doctoral Dissertation

Author: Ali Elminejad
Study program: Economics and Finance
Supervisor: Tomáš Havránek, Ph.D.
Year of defense: 2024

Abstract

This dissertation presents a comprehensive meta-analysis of deep parameters in economics, focusing on three significant areas: relative risk aversion, the Frisch elasticity of labor supply, and the Calvo parameter. Each study in this dissertation aims to refine our understanding of these parameters by synthesizing large amounts of empirical data, thereby addressing the pervasive issues of publication bias and estimation discrepancies prevalent in economic literature.

The first article in this dissertation undertakes a meta-analysis of the literature on relative risk aversion, employing the consumption Euler equation and distinguishing estimates from calibrations. Applying different techniques, the article corrects for publication bias and model uncertainty and reveals a divergence between estimated values in the economics and finance literature.

Moving to labor market dynamics, the second article addresses the Frisch elasticity of labor supply, a critical parameter for studying the response of the labor market to economic conditions or policy shifts. The analyses in the second article correct for publication bias and highlight the effect of identification bias on estimated elasticities at extensive and intensive margins.

Finally, the third article delves into estimating the Calvo parameter within the empirical New Keynesian Phillips Curve. The study identifies the distortion effects of publication bias and the impact of research design on reported estimates. The nuanced analysis underscores the sensitivity of the Calvo parameter to various modeling choices, such as the forcing variable and instruments. Hence, the findings offer insights for more accurate calibrations in modeling the New Keynesian Phillips Curve.

JEL Classification	C11, C83, D81, D90, E24, E31, J21
Keywords	Meta-analysis, publication bias, Bayesian model averaging, risk aversion, Frisch elasticity, Calvo
Title	Essays on the Meta-Analysis of Deep Parameters

Acknowledgments

I am grateful especially to my advisor, Prof. Tomáš Havránek, for his support throughout the doctoral program at Charles University. Without his advice and encouragement, this dissertation would not have been completed. I owe immense gratitude to my partner and wife, Nino, whose support and encouragement have been my anchor throughout my doctoral studies. Her patience and love made all the difference.

In the process of writing this dissertation, I benefited from the financial support provided by Charles University Grant Agency (GAUK) project No. 736120, and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 870245.

Typeset in LaTeX using the IES Thesis Template.

Bibliographic Record

Elminejad, Ali: *Essays on the Meta-Analysis of Deep Parameters*. Doctoral Dissertation. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 220. Advisor: Tomáš Havránek, Ph.D.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Deep parameters in economics are fundamental constants that describe underlying aspects of the economy. Since the argument proposed by Lucas (1976), widely known as *the Lucas critique*, these parameters become a critical component of macroeconomic modeling, particularly DSGE models. The Lucas critique posits that to forecast the outcomes of policy interventions accurately, it is essential to base our models on *deep parameters* that reflect individual behavior (e.g., preferences, technology, and resource constraints), collectively referred to as *microfoundations*. These parameters are considered *deep* because they are stable and invariant to changes in economic policy or other external conditions. They include factors like the rate of time preference, elasticity of substitution, production functions, and utility functions.

Deep parameters are crucial for calibrating economic models so that the models can accurately replicate observed phenomena. Summers (1991) emphasizes the role of deep parameters in understanding the primary driving forces behind economic decisions. Knowing them helps predict how economic environment changes (like tax changes or technological advancements) will alter consumer and producer behavior. Furthermore, Summers (1991) highlights the importance of deep parameters in long-term forecasting and simulations, stating that their stability allows economists to project future economic conditions

under different scenarios without the parameters themselves being influenced by short-term economic fluctuations.

Moreover, addressing comparative statics in economic theory, Estrella and Fuhrer (2003) show that comparing different equilibrium states as deep parameters change while keeping everything else constant helps in understanding how changes in fundamental aspects of the economy (like technology or preferences) might affect economic outcomes. They also discuss using deep parameters to make meaningful comparisons between countries. Differences in these parameters explain why similar policies have different outcomes in different countries or why some countries grow faster.

This dissertation focuses on three parameters used in macroeconomic modeling. All three articles in this dissertation are meta-analyses investigating sources of heterogeneity and bias in estimating deep parameters. Meta-analysis in economics is a powerful statistical tool that synthesizes results from multiple individual studies, aiming to understand the biases and trends across a body of economic research. This approach is particularly valuable in economics, where individual studies often have conflicting results or vary significantly in design, scope, and quality. All articles in this dissertation follow a homogenized set of tools and methods to fulfill the standards in conducting meta-analysis in line with Stanley et al. (2013) and Havránek et al. (2020).

I use advanced linear and nonlinear techniques to study the extent of publication bias in the literature. Furthermore, I employ Bayesian model averaging (BMA) as a natural solution to model uncertainty to scrutinize different aspects of the research framework in which the parameters are estimated. This method estimates many models that include various combinations of the collected explanatory variables and weights individual models by goodness of fit and parsimony. The BMA method is crucial for statistical analysis and decision-making where uncertainty across multiple model specifications is a significant concern.

Its ability to integrate multiple sources of evidence and its robustness to model misspecification make it a powerful tool in many domains.

The main findings show that in addition to various sources of heterogeneity, a significant publication bias is pervasive in the literature. The first article, *"Estimating Relative Risk Aversion from the Euler Equation: The Importance of Study Design and Publication Bias"*, co-authored with Tomáš Havránek and Zuzana Irsova, tackles the wide variation in estimates of relative risk aversion. Almost every structural model requires assumptions concerning relative risk aversion, and dozens of studies have estimated the corresponding coefficient using the consumption Euler equation. However, no consensus on the appropriate calibration values has emerged. Collecting 1,021 estimates from 92 studies that use the consumption Euler equation to measure relative risk aversion and that disentangle it from intertemporal substitution, we show that calibrations of risk aversion are systematically larger than estimates thereof. Moreover, reported estimates are systematically larger than the underlying risk aversion because of publication bias. After correcting the bias, the literature suggests a mean risk aversion of 1 in economics and 2–7 in finance contexts. The reported estimates are driven by the characteristics of data (frequency, dimension, country, stockholding) and utility (functional form, treatment of durables).

The second article, *"Intertemporal Substitution in Labor Supply: A Meta-Analysis addresses the Frisch elasticity of labor supply"*, co-authored with Tomáš Havránek, Roman Horvath, and Zuzana Irsova, is published in the Review of Economic Dynamics. This article addresses the crucial role of intertemporal substitution (Frisch) elasticity of labor supply in predicting how labor supply responds to changes in wages or tax policies. This article conducts two separate meta-analyses on both intensive and extensive margins. We show that the mean reported estimates of the elasticity are exaggerated due to publication bias. For both the intensive and extensive margins, the literature provides over

700 estimates, with a mean of around 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 negligible. A total hours elasticity of about 0.25 is the most consistent with empirical evidence. To investigate the differences in reported elasticities to differences in estimation context, we collect 23 additional variables reflecting the study design and employ BMA 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 a total hours elasticity above 0.5.

Finally, the third article, *"The Calvo Parameter Revisited: An Unbiased Insight"*, is a solo-authored paper published in the Applied Economics Letters. This study examines the sources of heterogeneity in the estimates of the Calvo parameter, integral to the New Keynesian Phillips Curve (NKPC), for modeling inflation dynamics. Conducting novel linear and nonlinear techniques on 509 estimates collected from 40 studies, I show how publication bias shifts reported estimates towards more conventional values used in model calibrations. Moreover, BMA results indicate that the reported estimates are systematically affected by various aspects of research design, particularly the choice of forcing variable in the NKPC, instrument selection, and authors' affiliation.

Together, these articles illuminate the complexities of estimating deep parameters in macroeconomics and highlight the essential role of meta-analysis in providing a clearer and more accurate picture. By systematically addressing issues like publication bias and research design variability, the results in this dissertation enhance our understanding of critical structural parameters. However, issues such as p-hacking and attenuation bias are not fully addressed in this dissertation and can be further explored in future research.

References

Estrella, A. and Fuhrer, J. C. (2003). Monetary policy shifts and the stability of monetary policy models. *Review of Economics and Statistics*, 85(1):94–104.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and Van Aert, R. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, 34(3):469–475.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland.

Stanley, T. D., Doucouliagos, H., Giles, M., Heckemeyer, J. H., Johnston, R. J., Laroche, P., Nelson, J. P., Paldam, M., Poot, J., Pugh, G., et al. (2013). Meta-analysis of economics research reporting guidelines. *Journal of economic surveys*, 27(2):390–394.

Summers, L. H. (1991). The scientific illusion in empirical macroeconomics. *The Scandinavian Journal of Economics*, pages 129–148.

Chapter 2

Estimating Relative Risk Aversion from the Euler Equation: The Importance of Study Design and Publication Bias

Abstract

Estimates of relative risk aversion vary widely, but no study has attempted to quantitatively trace the sources of the variation. We collect 1,021 estimates from 92 studies that use the consumption Euler equation to measure relative risk aversion and that disentangle it from intertemporal substitution. We show that calibrations of risk aversion are systematically larger than estimates thereof. Moreover, reported estimates are systematically larger than the underlying risk aversion because of publication bias. After correction for the bias, the literature suggests a mean risk aversion of 1 in economics and 2–7 in finance contexts. The reported estimates are driven by the characteristics of data (frequency, dimension, country, stockholding) and utility (functional form, treatment of durables). To obtain these results we use recently developed nonlinear techniques to correct for publication bias and Bayesian model averaging techniques to account for model uncertainty.

Keywords: Euler equation, risk aversion, Epstein-Zin, meta-analysis, publication bias, Bayesian model averaging

JEL Codes: C83, D81, D90

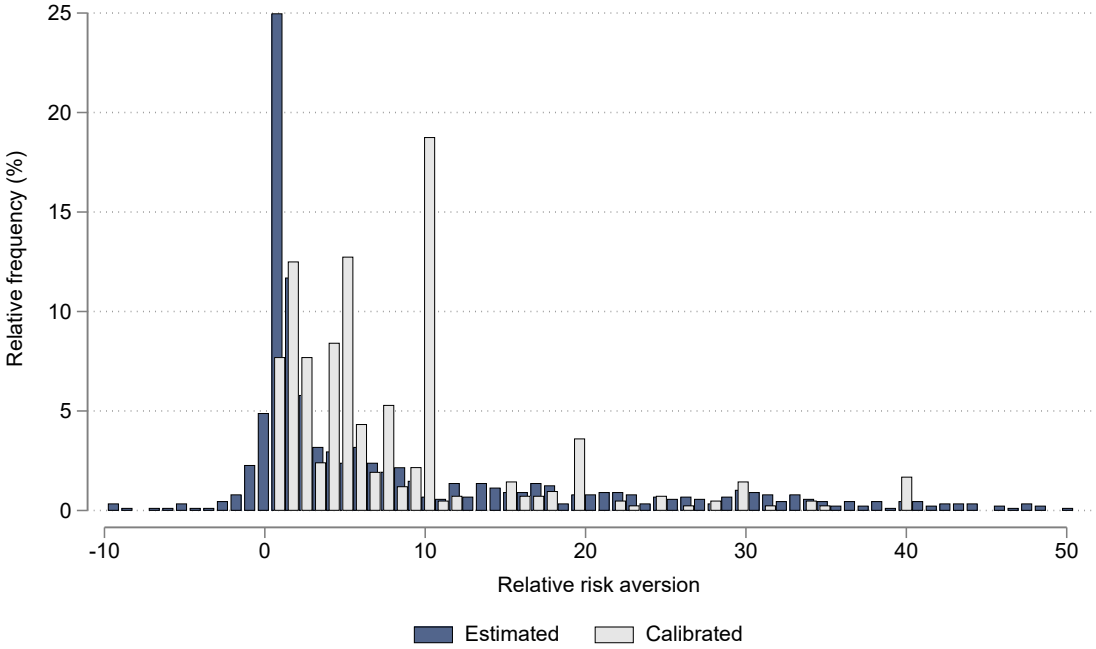
This paper is a joint work with Tomáš Havránek and Zuzana Irsova. An online appendix with data and code is available at `meta-analysis.cz/risk`.

2.1 Introduction

Risk aversion is a key concept in economics and finance. Almost every structural model requires assumptions concerning relative risk aversion, and dozens of studies have estimated the corresponding coefficient using the consumption Euler equation. Yet no consensus on the appropriate calibration values has emerged, as Figure 2.1 demonstrates: common values are 2.5, 5, and 10, but 1 and 20 also appear often. Remarkably, the distribution of calibrations does not match the distribution of estimates. The most common estimated value is 1, while the most common calibration is 10. But the figure also shows that almost every calibrated value up to at least 50 can be justified by some empirical estimates. There are few guidelines on the calibrations of relative risk aversion, and no quantitative synthesis (or meta-analysis) has attempted to shed light on the issue. That is what we attempt to deliver in this paper.

The absence of a meta-analysis on the topic can perhaps be explained by the sheer size of the literature on risk aversion. Risk aversion can be estimated using lab experiments, surveys, labor-supply behavior, auction behavior, choices in insurance contracts, option prices, and game show contestant behavior (see, for example, Zhang et al. 2014). We focus on the consumption Euler equation approach, which constitutes the benchmark framework employed in economics and finance. The problem is that most studies in this literature assume power utility, which means that relative risk aversion equals the reciprocal of the elasticity of intertemporal substitution, and hence the interpretation of the estimated parameter is unclear. We thus concentrate on the subset of the literature that separates risk aversion from intertemporal substitution. The separation is typically done by employing Epstein-Zin preferences (Epstein and Zin 1989; 1991), but can also be achieved using habits in consumption, expected utility with a reference level of consumption, ambiguity aversion, or disappointment

Figure 2.1: Calibrations of risk aversion overtop most estimates thereof



Notes: The figure shows histograms of i) 1,021 estimates or relative risk aversion collected from 92 studies and ii) 446 calibrations of relative risk aversion collected from 200 studies. In both cases we only consider studies that separate risk aversion from intertemporal substitution. For ease of exposition, values below −10 and above 50 are excluded from the figure but included in all statistical tests. Summary statistics are available in Table 2.1. Separate figures for economics and finance literatures are available in Figure 2.B1 and Figure 2.B2, respectively.

aversion. Even this subset of the Euler equation literature yields 1,021 estimates from 92 studies. To construct Figure 2.1 we also collect 446 calibrations from 200 studies, once again only those that break the link between risk aversion and intertemporal substitution.

Four previous studies are intimately related to the analysis we present. Havranek (2015) conducts a meta-analysis of the elasticity of intertemporal substitution in consumption. After correcting the literature for various biases, he argues that the best guess concerning the mean elasticity of substitution is 1/3. Because almost all studies in his sample use power utility, the finding translates to the relative risk aversion of 3—if we accept the argument by Kocherlakota (1990), contrary to Hall (1988), that the parameter derived from

the corresponding Euler equation is more informative about risk aversion than intertemporal substitution. Ascari et al. (2021) present a recent and meticulous estimation, robust to weak instruments, of all parameters that can be derived from the consumption Euler equation. They find that the potential range for relative risk aversion is wide. Brown et al. (2023) conduct a meta-analysis of loss aversion, a concept related to but distinct from relative risk aversion as commonly used in economics, and find that the mean loss aversion is around 2 after correction for several biases. Imai et al. (2021) present a meta-analysis of the present bias, which some argue (prominently, Dean and Ortoleva 2019) is strongly related to risk preferences. The corrected mean present bias recovered by Imai et al. (2021) is between 0.95 and 0.97.

Key issues for meta-analysis are the twin problems of publication bias and p-hacking. Publication bias describes a situation in which authors, referees, or editors, intentionally or not, refuse to publish estimates that are statistically insignificant or inconsistent with the theory (for example, have the wrong sign). P-hacking is the effort by authors, again intentional or not, to produce publishable results: for example, by trying different subsamples or control variables until the estimate reaches statistical significance. McCloskey and Ziliak (2019) invoke a nice analogy to the Lombard effect in psychoacoustics: speakers involuntarily increase their vocal effort in the presence of noise. In a similar way can researchers respond to noise in their data or techniques and try harder till they obtain a point estimate large enough to compensate for the large standard error. Note that publication bias and p-hacking are observationally equivalent, so for parsimony we will use the term publication bias to describe both, as is common in the meta-analysis literature. Many studies have recently discussed how publication bias can exaggerate empirical estimates in economics (Brodeur et al. 2016; Bruns and Ioannidis 2016; Card et al. 2018; Christensen and Miguel 2018; DellaVigna et al. 2019; Blanco-Perez and Brodeur 2020; Brodeur et al.

2020; Ugur et al. 2020; Xue et al. 2020; Neisser 2021; Stanley et al. 2021; DellaVigna and Linos 2022; Stanley et al. 2022), and the exaggeration can be twofold or more (Ioannidis et al. 2017). Publication bias is natural, common in economics, and does not imply cheating or any ulterior motives on the part of the researchers. But it is a serious problem for the interpretation of the results in the literature, a problem meta-analysis can tackle.

Most meta-analysis techniques used for publication bias correction in economics and finance rely on the Lombard effect and regress estimates on their standard errors (meta-regression). Evidence of a nonzero slope is commonly taken as evidence for publication bias, and the constant in the regression measures the mean estimate conditional on maximum precision, often interpreted as the mean corrected for the bias. There are two problems with such a strategy. First, as shown by Andrews and Kasy (2019) and Stanley and Doucouliagos (2014), publication bias can be a nonlinear function of the standard error. Second, as discussed by Havranek et al. (2023), the assumption of no correlation between estimates and standard errors in the absence of publication bias can be problematic because of unobserved heterogeneity that affects both estimates and standard errors. To address these two problems, we employ recently developed nonlinear tests for publication bias: the selection model by Andrews and Kasy (2019), the weighted average of adequately powered estimates (Ioannidis et al. 2017), the stem-based technique (Furukawa 2021), the endogenous kink model (Bom and Rachinger 2019), and the p-uniform* technique (van Aert and van Assen 2021).

In the second part of the analysis we investigate the heterogeneity in the reported estimates of relative risk aversion. We identify 30 characteristics of data, specification, estimation, and publication that reflect the context in which the estimates are obtained and that may affect the estimates. The characteristics are so numerous because of the many choices researchers have to make when

specifying their models. In consequence, substantial model uncertainty arises in meta-analysis when we want to relate estimates of risk aversion to estimation context. As a solution we use Bayesian model averaging (see, e.g., Zeugner and Feldkircher 2015; Steel 2020), which is the natural response to model uncertainty in a Bayesian setting; moreover, it is computationally less cumbersome than frequentist alternatives. Bayesian model averaging also allows us to partially address collinearity by employing the dilution prior (George 2010), which penalizes models with a small determinant of the correlation matrix.

We find substantial publication bias in the empirical literature on relative risk aversion. The mean amount of exaggeration due to the bias is striking: about seven-fold in both economics and finance. The corrected mean relative risk aversion is 1 in the economics literature and 2–7 in the finance literature (where different correction techniques give quantitatively different results, but all agree that publication bias is strong). The correction for publication bias further widens the gap between typical estimates and typical calibrations presented earlier in Figure 2.1. In particular, the value of 10 most frequently used for calibration is inconsistent with the bulk of empirical estimates. In contrast, the second most common calibration, 5, is well within the plausible range of estimates suggested by the literature in finance (but not economics) contexts. Note also that the mean estimate of 1 obtained for economics does not lend itself to the recommendation of the logarithmic utility function in that field. The reason is, as we have mentioned earlier, that the elasticity of intertemporal substitution is typically not 1 but around 1/3 (Havranek 2015). In finance contexts, power utility with relative risk aversion set at 3 thus seems relatively consistent with empirical evidence.

When we allow for heterogeneity by employing Bayesian model averaging, we confirm the finding of strong publication bias and a substantial difference in estimated risk aversion between economics and finance contexts—even after

other aspects of data and methods are controlled for. In addition, studies that focus on stockholders tend to find substantially smaller values of risk aversion, which is consistent with both intuition and previous results (such as Mankiw and Zeldes 1991). Finally, reported estimates of relative risk aversion are systematically related to data characteristics (frequency, dimension, and country coverage) and the definition of the utility function (the assumption of separability between durables and nondurables and the use of Epstein-Zin preferences in contrast to other methods for separating risk aversion from intertemporal substitution). The results are reasonably robust to alternative priors for Bayesian model averaging.

2.2 Data

Details on the estimation of relative risk aversion in the context of the consumption Euler equation are available in Section 2.B; the estimation approaches followed by most studies are also clearly described by Epstein and Zin (1991) and Vissing-Jørgensen and Attanasio (2003). A more general overview of modeling risk aversion is presented by O'Donoghue and Somerville (2018). Section 2.A provides details on the way we search the literature for estimates of relative risk aversion. We start with a search query in Google Scholar, which we prefer over alternative databases because of its universal coverage and full-text capabilities. The search query yields more than 3,500 studies. For feasibility, we only inspect the first 1,500 studies returned by the search. We read the abstracts of these studies and download those that indicate any chance of containing empirical estimates of risk aversion (about a half of the examined studies).

We read the downloaded studies and include those that conform to the following three criteria. First, the study must use the consumption Euler equation to obtain an empirical estimate of the coefficient of relative risk aversion.

Second, the estimate must be reported together with the corresponding standard error or any statistics from which the standard error can be computed. Third, the study must separate risk aversion from intertemporal substitution. We collect both published and unpublished papers, and terminate the search on May 16, 2022. The search yields 92 papers (called "primary studies" in the meta-analysis terminology and listed in Table 2.B1), which together provide 1,021 estimates of relative risk aversion. The sample of calibration studies is assembled using a similar search strategy with the following differences: in the search query we replace the word "estimate" with "calibration", restrict our attention to published papers, and stop once we collect 200 usable studies (ranked by the order they appear in the Google Scholar reply to our query). This approach yields 446 individual calibrated values of relative risk aversion.

In addition to calibrations, estimates, and the estimates' standard errors, we also collect 30 variables, described in Section 2.4, that reflect the context in which the estimates are obtained in primary studies: the characteristics of data, specification, estimation, and publication. This means we collect manually more than 30,000 data points. To reduce the danger of mistakes and typos, two of the co-authors collect the data independently, and the third co-author resolves inconsistencies between these two datasets. The resulting clean dataset is available in the online appendix at `meta-analysis.cz/risk` together with the code used in this analysis and the list of 200 calibration studies.

Throughout the paper we distinguish between estimates obtained in economics and finance contexts. The precise boundary is hard to draw: estimates in economics are often, but not always, derived from approaches that focus on the entire economy, while finance estimates tend to focus almost exclusively on asset prices (see Section 2.B for details). We choose a classification based on the journal in which the primary or calibration study is published and follow the categories defined by the Web of Science. If in the Web of Science the journal

Table 2.1: Summary statistics of estimated and calibrated relative risk aversion

Panel A: Estimates

	Observations	Mean	Median	Standard deviation
All 92 studies	1,021	23.36	3.77	98.58
Economics (58 studies)	590	7.50	1.26	30.74
Finance (34 studies)	431	45.05	17.82	144.71

Panel B: Calibrations

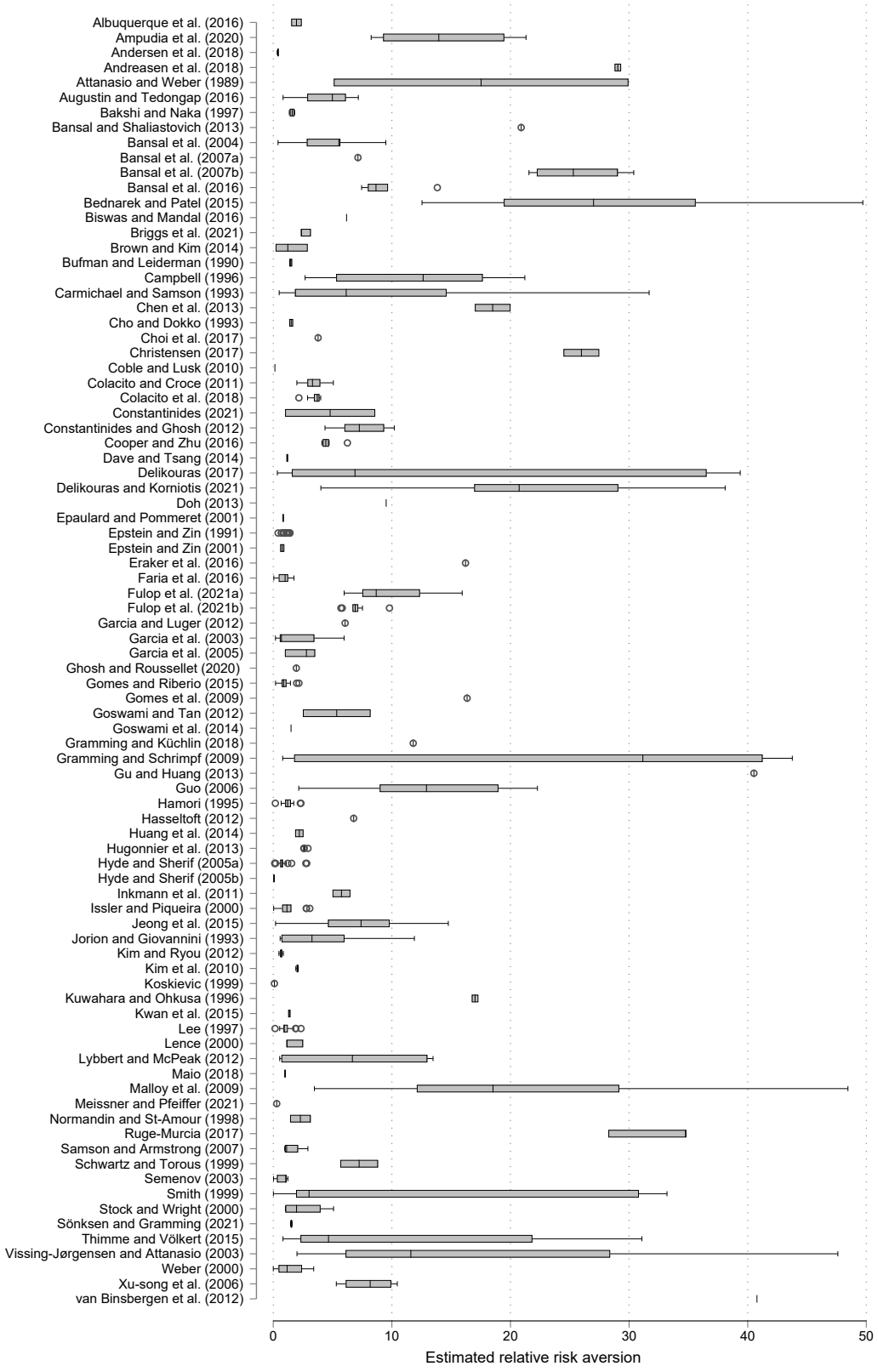
	Observations	Mean	Median	Standard deviation
All 200 studies	446	14.33	6.00	30.10
Economics (115 studies)	237	17.14	6.00	35.74
Finance (85 studies)	209	11.12	6.00	21.66

Notes: We only consider studies that separate risk aversion from intertemporal substitution. Studies are classified into economics and finance categories based on the journals they were published in and using the journal classification of the Web of Science. If in the Web of Science the journal is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. If a study is unpublished (15 studies in total), we classify it based on the prevailing publications of the corresponding author. In the meta-analysis we winsorize estimates at the 5% level. Summary statistics for benchmark calibrations from each study in Panel B are reported in Table 2.B2.

is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. If a study is unpublished (15 primary studies in total), we classify it based on the prevailing publications of the corresponding author. In such a way each study can be unambiguously classified into either economics or finance.

Table 2.1 presents the summary statistics of estimates and calibrations, and Figure 2.2 shows that the estimates vary within and across studies. As we have noted in the discussion of Figure 2.1, calibrations of risk aversion in the literature tend to be larger than most empirical estimates. But Table 2.1 shows a different story between economics and finance. In economics, calibrations are indeed much larger than estimates, both in terms of mean and median values; the corresponding histogram is available in Figure 2.B1 in Section 2.B. In finance, the opposite is the case: estimates overtop calibrations (Figure 2.B2). Calibrations in both fields are very similar to each other, with a median of 6

Figure 2.2: Estimates of risk aversion vary both across and within studies



Notes: The length of each box represents the interquartile range (P25-P75), and the dividing line inside the box is the median value. The whiskers represent the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. For ease of exposition, outliers are excluded from the figure but included in all statistical tests.

and mean around 15 (The pattern holds for the set of benchmark calibrations from each study; see Table 2.B2.) Figure 2.B2 shows that while even in finance the estimates of risk aversion between 1 and 10 are the most common, values around 20 and larger are also routinely reported.

Curiously, therefore, calibrations of relative risk aversion in both fields seem to have little basis in the distribution of the empirical estimates of the parameter in a given field. Instead, many calibrations simply quote Mehra and Prescott (1985), who argue that 10 is a reasonable upper bound for the coefficient of relative risk aversion. Because large risk aversion is often sought for calibration (for example, to help explain the equity premium puzzle), it follows that 10 is the most frequently used calibration value by a large margin. Values of 2.5, 5, and 20 present the most common robustness checks to the baseline calibration. Our goal in this paper is to help reconnect calibrations of risk aversion to empirical estimates thereof, and the first necessary step is the correction of the estimates for publication selection bias.

2.3 Publication Bias

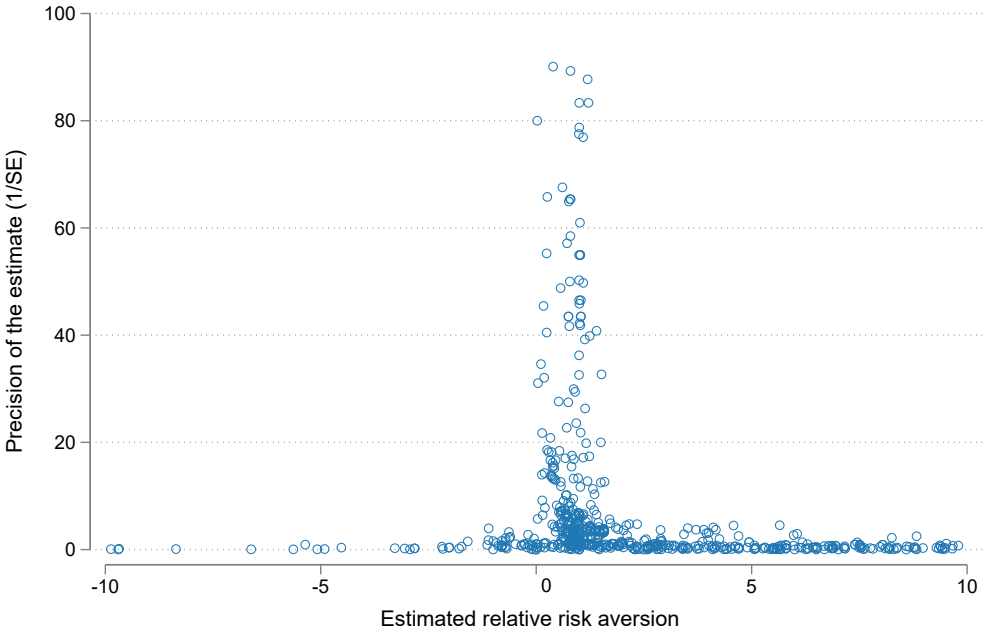
Economists expect that most people are risk-averse, and hence that the mean coefficient of relative risk aversion in any group is positive. This belief is reflected by the 446 calibrations shown earlier in Figure 2.1: all of them are positive. Negative or zero risk aversion bodes well with few economics and finance models. Of course, the underlying mean coefficient of relative risk aversion is most likely substantially positive. But unless it is huge, researchers will sometimes run into estimation contexts in which the estimate of the coefficient turns out to be insignificantly different from zero or even negative. Noise in the data or methods will produce such counter-intuitive results from time to time. In a similar way, noise will also produce estimates that are too large and

away from the true mean. The problem is that while it is difficult to identify the implausibly large estimates (no upper threshold exists for risk aversion), researchers immediately spot and investigate those that are negative or statistically insignificant. Given such unintuitive results, researchers may choose not to report them, or try a different specification in the hope of obtaining results that are consistent with their priors. Such a censoring drives the mean reported risk aversion upwards from the true value, and this is what meta-analysts call publication bias (Card and Krueger 1995; Stanley 2001).

The process leading to publication bias is not necessarily detrimental to science, and certainly it does not need to involve any ulterior motives on the part of the researchers. In most cases it will improve the inference of any study if it does not focus on negative or insignificant estimates of relative risk aversion. After all, these "nonsensical" estimates are likely to be caused by some problems in data or methods. But researchers do not winsorize: they typically treat small estimates with suspicion, but not those that are large. Thus, at the level of the entire literature, a bias arises that exaggerates the true mean effect. The tension between these two aspects of publication bias is nicely illustrated by the following quote due to Uhlig (2012, p. 38) about empirical evidence on monetary policy transmission:

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

Figure 2.3: The funnel plot suggests publication bias



Notes: In the absence of publication bias (and any small-sample and heterogeneity-related biases), the plot should form a a symmetrical inverted funnel. Outliers are excluded from the figure for ease of exposition but included in all tests.

How to test and correct for publication bias? The histogram of the estimates shown in Figure 2.1 does not really help, though it suggests that the bias is not universal: some negative estimates of risk aversion do appear in the literature. A neat way to measure publication bias is to compare the results of original studies and pre-registered replications (Kvarven et al. 2020), the latter being unlikely to suffer from much bias. But there are no pre-registered replications of studies estimating relative risk aversion from the Euler equation; in general, pre-registration is most efficient in the experimental literature where researchers cannot inspect their data prior to pre-registration (Olken 2015). To correct for the bias, we thus rely on techniques traditionally used by medical researchers and new methods recently developed by econometricians and psychologists.

The starting point is a visual examination of the so-called funnel plot, often used in medical research (Egger et al. 1997; Stanley and Doucouliagos 2010).

The funnel plot, Figure 2.3, is a scatter plot of point estimates on the horizontal axis and the estimates' precision (reciprocal of the standard error) on the vertical axis. In the absence of systematic heterogeneity, which will be examined in the next section, the most precise estimates should be close to the underlying mean coefficient of relative risk aversion. As precision decreases, the estimates should be more widely dispersed around the true mean value. Because in the absence of publication bias all estimates have the same chance of being reported, the funnel will be symmetrical: all imprecise estimates are published, both those that are negative and those that are huge and positive. Figure 2.3 shows that, first, the funnel is asymmetrical, which indicates publication bias against small estimates of risk aversion. Second, the most precise estimates are concentrated around 1.

Table 2.2 shows the results of more formal tests of funnel asymmetry and the underlying risk aversion beyond publication bias. The tests are regressions of estimates on standard errors and can also be interpreted as tests of the Lombard effect discussed in the Introduction (researchers increase their specification search effort in response to noise in their data or methods). The estimated slope in the regression measures the extent of publication bias. The intercept can be interpreted as the mean coefficient of relative risk aversion corrected for publication bias: if we assume that publication bias is indeed a linear function of the standard error. (This is a strong assumption that we will later relax.) We account for the obvious heteroskedasticity by weighting the regressions by inverse variance (Stanley and Doucouliagos 2014; 2015). We employ four specifications: standard weighted least squares, study-level fixed effects, study-level between effects, and a specification that additionally weights estimates by the inverse of the number of estimates reported by each study, thus giving each study the same weight. All specifications except between effects report standard errors clustered at the study level; for the first and last

Table 2.2: Funnel asymmetry tests indicate modest risk aversion beyond publication bias

Panel A: All studies				
	WLS	FE	BE	Study
Standard error	1.865***	2.287***	2.837	3.062***
(*publication bias*)	(0.362)	(0.713)	(1.760)	(0.893)
	[0.956, 2.577]			[1.251, 4.900]
Constant	1.199***	1.084***	1.590***	1.533***
(*mean corrected RRA*)	(0.257)	(0.194)	(0.235)	(0.412)
	[0.725, 2.130]			[0.673, 2.476]
Observations	1,021	1,021	1,021	1,021
Studies	92	92	92	92
Panel B: Economics				
	WLS	FE	BE	Study
Standard error	1.392***	1.411	4.119***	3.604***
(*publication bias*)	(0.540)	(1.146)	(1.361)	(0.827)
	[0.383, 2.506]			[2.007, 5.293]
Constant	1.085***	1.082***	0.714***	0.822***
(*mean corrected RRA*)	(0.261)	(0.211)	(0.178)	(0.243)
	[0.654, 2.059]			[0.351, 1.464]
Observations	590	590	590	590
Studies	58	58	58	58
Panel C: Finance				
	WLS	FE	BE	Study
Standard error	1.859***	3.476***	0.817	2.168
(*publication bias*)	(0.449)	(0.169)	(3.061)	(1.654)
	[0.050, 2.895]			[-1.197, 5.548]
Constant	2.390***	1.107***	3.223***	2.888***
(*mean corrected RRA*)	(0.675)	(0.134)	(0.423)	(0.732)
	[0.812, 4.006]			[1.062, 4.89]
Observations	431	431	431	431
Studies	34	34	34	34

Notes: We regress estimates of relative risk aversion on their standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are reported in parentheses. RRA = relative risk aversion. WLS = standard weighted least squares. FE = study fixed effects. BE = study between effects. Study = the inverse of the number of estimates reported per study is used as an additional weight. In square brackets we show the 95% confidence interval from wild bootstrap (Roodman et al. 2019). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

specification we also report confidence intervals based on wild bootstrap.

In all cases we obtain estimated coefficients for publication bias that are positive and large, in line with the funnel plot. Most of them are also statistically significant at the 5% level. Given that this test for publication bias is known to have relatively low power (Stanley 2008), the results are consistent with substantial bias. The corrected mean coefficient of relative risk aversion is around 1 for economics and 1–3 for finance, compared with uncorrected means of 7.5 and 45, respectively. The estimated exaggeration due to publication bias is striking and much larger than what is typical in economics: Ioannidis et al. (2017) report that the mean exaggeration due to publication bias is twofold. Next, we relax the assumption that publication bias is a linear function of the standard error, which has been criticized by Andrews and Kasy (2019) and Stanley and Doucouliagos (2014). In doing so, we rely on recently developed nonlinear models of publication bias.

We use five nonlinear techniques for publication bias correction. First, the weighted average of adequately powered estimates by Ioannidis et al. (2017). The technique estimates retrospective power for all estimates and yields a result that is the average of the estimates with power above 80% (weighted by inverse variance). Second, the selection model by Andrews and Kasy (2019). This rigorously founded technique estimates the probability that negative and insignificant estimates are not reported; the probability is then used to upweight these estimates. Third, the stem-based technique by Furukawa (2021). The technique exploits the trade-off between bias and variance: when more imprecise studies are added, publication bias increases, but variance decreases because more estimates are available. Furukawa (2021) minimizes the corresponding mean squared error that is the sum of bias and variance. Fourth, the endogeneous kink model by Bom and Rachinger (2019). The technique assumes that the relationship between estimates and standard errors is linear

Table 2.3: Nonlinear corrections for publication bias

Panel A: All studies	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Furukawa (2021)	Bom and Rachinger (2019)	van Aert and van Assen (2021)
Mean corrected RRA	1.318***	0.960***	1.467***	1.199***	0.367***
	(0.250)	(0.035)	(0.951)	(0.046)	[0.002]
Observations	1,021	1,021	1,021	1,021	1,021
Studies	92	92	92	92	92
Panel B: Economics	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Furukawa (2021)	Bom and Rachinger (2019)	van Aert and van Assen (2021)
Mean corrected RRA	1.172***	0.910***	0.474***	1.085***	0.366***
	(0.250)	(0.030)	(0.390)	(0.052)	[0.002]
Observations	590	590	590	590	590
Studies	58	58	58	58	58
Panel C: Finance	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Furukawa (2021)	Bom and Rachinger (2019)	van Aert and van Assen (2021)
Mean corrected RRA	2.535***	11.196***	6.100***	2.390***	0.625***
	(0.662)	(1.212)	(0.885)	(0.112)	[0.008]
Observations	431	431	431	431	431
Studies	34	34	34	34	34

Notes: RRA = relative risk aversion. Standard errors are reported in parentheses; the p-uniform* technique due to van Aert and van Assen (2021) only yields p-values, which we report in square brackets.$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

when precision is low but that no relationship exists when precision is sufficiently high. For example, if the p-value is 0.001, publication probability is not affected by small changes in the standard error. Fifth, the p-uniform* model by van Aert and van Assen (2021). The technique, developed in psychology, works with the distribution of p-values and uses the statistical principle that the distribution should be uniform at the true mean value of the coefficient of relative risk aversion.

The first four techniques introduced above assume that there is no correlation between estimates and standard errors in the absence of publication bias. This is a common meta-analysis assumption that traces its roots back to medical research, where meta-analysis was first developed and applied. But in economics the assumption is problematic. Most of the research here is observational, which means much more heterogeneity and choices that researchers

have to make: and not all of the choices are (or can be) reported. It is entirely plausible that certain aspects of data or methods affect both estimates and standard errors, thereby creating a correlation even in the absence of publication bias. For example, studies using instrumental variables may correct for an endogeneity bias, but the resulting estimates also tend to be less precise. Indeed, the estimates are correlated with standard errors even if we employ a subsample of estimates that are likely to be published in any case because they are highly statistically significant (with p-values < 0.005); see Table 2.C3 in Section 2.C. In Table 2.C2 we perform a test, due to Kranz and Putz (2022), of the Andrews and Kasy (2019) model. The lack of correlation between estimates and standard errors in the absence of publication bias is the key assumption of the model, but the Kranz and Putz (2022) test concerns the Andrews and Kasy (2019) model as a whole, including the assumption of constant publication probabilities for estimates with the same classification of statistical significance (for example, p-values between 0.05 and 0.1). The test rejects the validity of the model in the case of relative risk aversion. Only the p-uniform* does not rely on the uncorrelation assumption because here the identification is based on p-values, not on estimates and their standard errors.

The results of the nonlinear tests are shown in Table 2.3. All tests corroborate strong publication bias: the corrected mean coefficients of relative risk aversion are always much smaller than uncorrected means shown earlier in Table 2.1. But the individual results vary. The p-uniform* technique, which can be considered conceptually superior to other models in the context of risk aversion since it does not rely on the uncorrelation assumption, yields values of risk aversion below 1 for both economics and finance. The selection model yields a large estimate for finance, 11, but we have seen that in our dataset the model probably does not work well. The remaining results are more consistent and suggest relative risk aversion around 1 in economics and 2–6 in finance:

with the qualification that our interpretation of the strength of publication bias is conservative because p-uniform* suggests an even stronger exaggeration. Finally, we also apply two new tests of p-hacking by Elliott et al. (2022) in Table 2.C1, Section 2.C. These tests also do not rely on the uncorrelation assumption, but need a huge sample and only test p-hacking without estimating the corrected risk aversion. Using these tests we reject the hypothesis of no bias in the entire sample but not in the individual subsamples of economics and finance studies.

2.4 Heterogeneity

Another way to relax the uncorrelation assumption is to explicitly allow for heterogeneity among the estimates of relative risk aversion. To this end we collect 30 aspects of the context in which the estimates are obtained. Using these additional variables we seek answers to three questions: Are our findings regarding publication bias robust to heterogeneity? Do some aspects of data or methods affect the reported estimates systematically? What is the literature's best guess regarding relative risk aversion in various contexts after correction for publication bias?

The variables are summarized in Table 2.4 and discussed in detail in the Appendix, Subsection 2.D.1. For ease of exposition we divide them into four groups: data characteristics, specification characteristics, estimation techniques, and publication characteristics. The list of variables that control for the context in risk aversion estimation is potentially unlimited, but we do our best to account for differences that are most commonly discussed in the literature. Figure 2.D1 shows that even with so many variables, collinearity is likely not a major issue for our analysis. Even so, we employ techniques that take collinearity into account.

Table 2.4: Definition and summary statistics of explanatory variables

Variable	Description	Mean	SD
Standard error	The standard error of the estimated coefficient of relative risk aversion.	76.65	730.63
Data characteristics			
Time span	The logarithm of the time span of the data used to estimate RRA.	3.45	0.92
Midpoint	The logarithm of the median year of the data used minus the earliest median year observed in primary studies.	3.82	0.63
Panel	= 1 if panel data are used (reference category: time series).	0.04	0.19
Cross-section	= 1 if cross-sectional data are used (reference category: time series).	0.20	0.40
Monthly	= 1 if data frequency is monthly or higher (reference category: annual).	0.25	0.43
Quarterly	= 1 if data frequency is quarterly (reference category: annual).	0.50	0.50
US	= 1 if the estimate relates to the United States (reference category: other countries).	0.74	0.44
EU	= 1 if the estimate relates to European countries (reference category: other countries).	0.11	0.31
Asia	= 1 if the estimate relates to developed Asian countries (reference category: other countries).	0.03	0.18
Developing	= 1 if the estimate relates to developing countries, including China (reference category: other countries).	0.06	0.24
Specification characteristics			
Epstein-Zin	= 1 if preferences are of the Epstein-Zin type (the remaining estimates are derived from specifications with internal habits, expected utility with a reference level of consumption, ambiguity aversion, or disappointment aversion).	0.90	0.30
Long-run risk	= 1 if estimation features long-run risks.	0.32	0.47
Fixed EIS	= 1 if the value of the elasticity of intertemporal substitution is fixed when estimating RRA.	0.25	0.43
Nonseparable durables	= 1 if the model allows for nonseparability between durable and nondurable consumption.	0.13	0.33
Total consumption	= 1 if total consumption is used instead of nondurable consumption.	0.10	0.30
Exact Euler	= 1 if the exact Euler equation is estimated instead of the log-linearized one.	0.37	0.48
Human capital	= 1 if human capital is accounted for in the estimation.	0.10	0.30
Stockholder	= 1 if the estimate relates to stockholders or wealthy households (reference category: mixed sample).	0.12	0.32
Nonstockholder	= 1 if the estimate relates to nonstockholders or poor households (reference category: mixed sample).	0.05	0.21
Estimation techniques			
Experimental	= 1 if the estimate is based on (quasi-)experimental data.	0.02	0.15
Implied	= 1 if the value of RRA is not reported explicitly but can be computed from other reported parameters.	0.12	0.32
GMM	= 1 if the generalized method of moments is used (reference category: OLS).	0.59	0.49
Simulations	= 1 if nonparametric simulation-based methods are used (reference category: OLS).	0.17	0.37
Second lag	= 1 if only second or higher lags are included among instruments.	0.16	0.36
Market return included	= 1 if market return is included among instruments.	0.32	0.47
Consumption included	= 1 if consumption is included among instruments.	0.35	0.48
Publication characteristics			
Publication year	The logarithm of the year when the study first appeared in Google Scholar minus the year when the earliest study in our dataset appeared in Google Scholar.	2.84	0.63
Top journal	= 1 if the estimate comes from a study published in the top five economics or top three finance journals.	0.30	0.46
Finance journal	= 1 if the estimate is reported in a finance journal.	0.42	0.49
Citations	The logarithm of the number of per-year citations of the study, according to Google Scholar.	1.72	1.40

Notes: All estimates that we collect are derived from specifications that separate risk aversion from intertemporal substitution. RRA = relative risk aversion; EIS = elasticity of intertemporal substitution; GMM = general method of moments; SD = standard deviation. The table excludes the definition and summary statistics of reference categories, which are omitted from Bayesian model averaging. Regarding the variable *Finance journal*, we use the classification of the Web of Science. If in the Web of Science the journal is included in both economics and finance categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking.

Because we have so many variables, we need to use methods that account for model uncertainty. While all of the variables we collect have been implicated in the literature to potentially affect the reported risk aversion, it is unclear whether all variables indeed belong to the best model. If not, then the effects of important variables will be imprecisely estimated, perhaps drastically so. A natural solution to model uncertainty arises in the Bayesian framework as Bayesian model averaging (see Steel 2020, for a great overview). Bayesian model averaging estimates many models that include various combinations of the explanatory variables we have collected and weights individual models by goodness of fit and parsimony. Because in our case there are too many possible models, we simplify this computationally demanding task by employing the Metropolis-Hastings algorithm of the `bms` package for R by Zeugner and Feldkircher (2015), which walks only through the most likely models. We also employ the dilution prior (George 2010), which accounts for collinearity by adding a weight that is proportional to the determinant of the correlation matrix of the variables included in the individual model. Unfortunately Bayesian model averaging can only be combined with the linear test of publication bias, but we have shown in the previous section that the results of the linear tests are broadly consistent with more advanced nonlinear techniques.

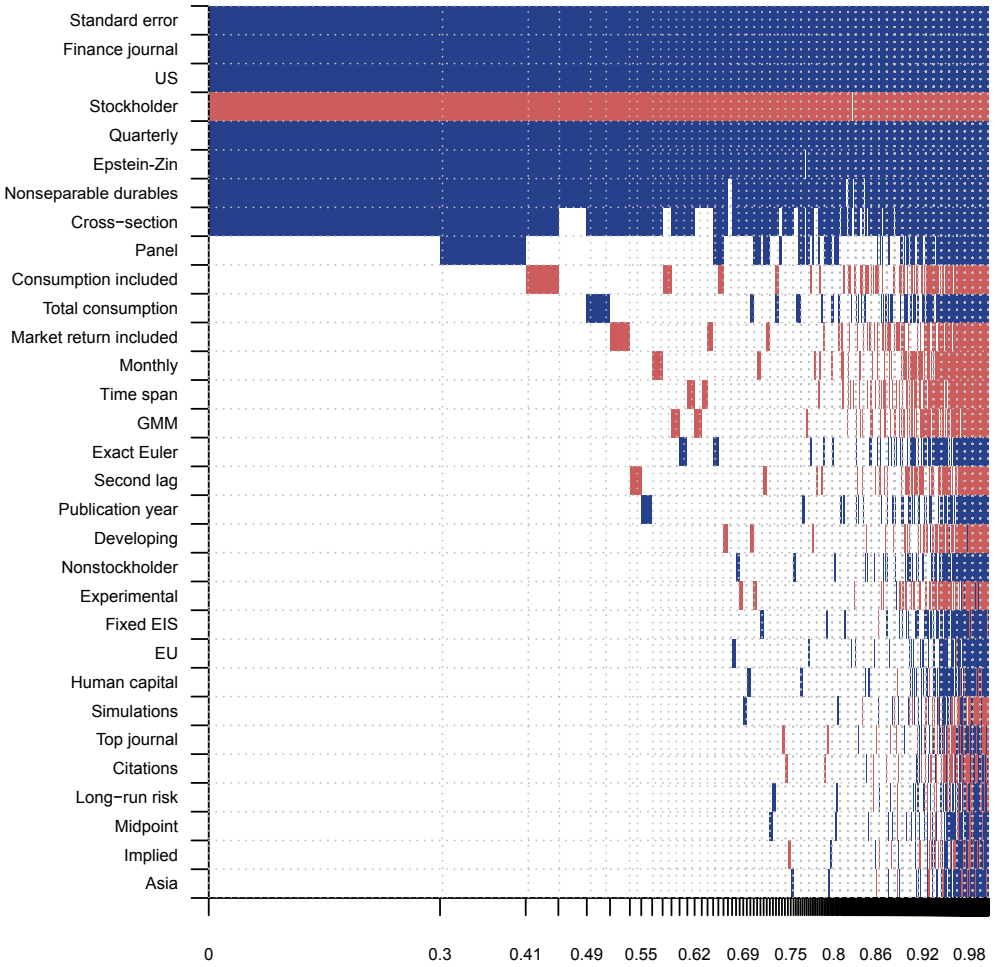
The results of Bayesian model averaging are summarized graphically in Figure 2.4; more details are available in Table 2.D1 and Figure 2.D2 in Subsection 2.D.2. The horizontal axis denotes cumulative posterior model probabilities: the weights received by each model. The most informative individual models, denoted by columns, therefore, are depicted on the left. Variables are sorted by posterior inclusion probability (the sum of posterior model probabilities of all models in which the variable is included) in descending order. This ordering means that the variables most useful in explaining the variation in estimated risk aversion are depicted at the top of the figure. The single most

important variable is the standard error, which corroborates our previous results concerning publication bias. In total, there are 8 variables with posterior inclusion probability above 0.5, which means that these variables are systematically related to the published coefficients of relative risk aversion. The results of Bayesian model averaging can be sensitive to the priors used, but Figure 2.5 and Table 2.D2 show that posterior inclusion probabilities do not change much when we apply alternative priors sometimes used in the literature.

The numerical results of Bayesian model averaging are reported in the left-hand part of Table 2.5. The right-hand part shows a simple frequentist robustness check, in which we run ordinary least squares using only the variables with posterior inclusion probability above 0.5 in Bayesian model averaging. The robustness check is broadly consistent with the results of Bayesian model averaging, but finds borderline statistical significance for several of the variables. The point estimates, however, are similar and suggest large effects of these characteristics. We find that, even if we control for estimation context, finance journals tend to report coefficients of relative risk aversion substantially larger than economics journals: by about 6. Another intuitive result is that stockholders are less risk-averse than nonstockholders. Again the difference in relative risk aversion is about 6. Next, we find that the results are driven by data and estimation characteristics: data dimension (cross-section vs. time series vs. panel data), data frequency (monthly vs. quarterly vs. annual), regional coverage (US vs. other countries), the specification of the utility function (Epstein-Zin vs. other approaches), and treatment of durables (separability vs. nonseparability). The heterogeneity results are described in more detail in Subsection 2.D.2.

Finally, we compute relative risk aversion implied by the literature for different settings after correction for publication bias and other potential biases. For this exercise we use the results of Bayesian model averaging and compute

Figure 2.4: Model inclusion in Bayesian model averaging



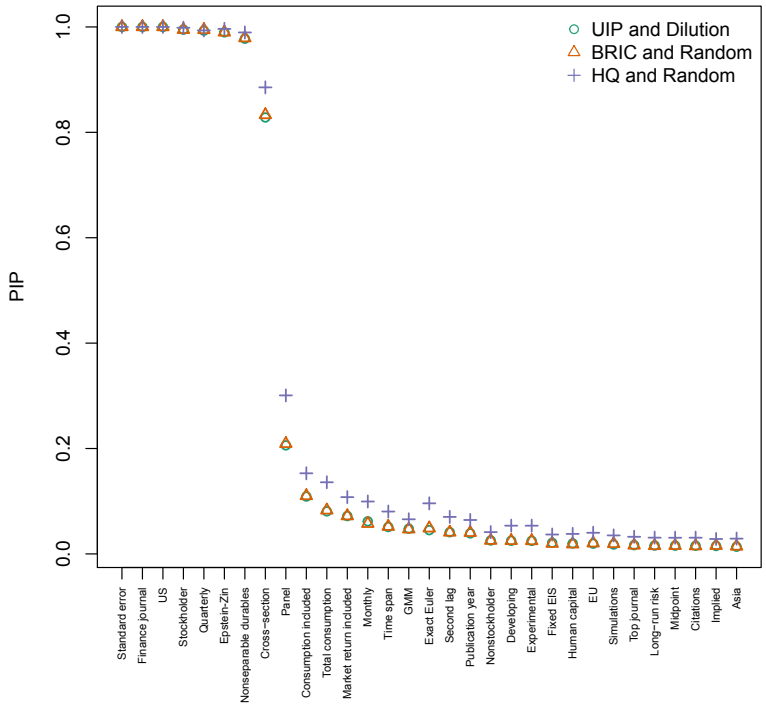
Notes: The response variable is the reported estimate of relative risk aversion; all estimates that we collect are derived from specifications that separate risk aversion from intertemporal substitution. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes cumulative posterior model probabilities. The estimation is based on the agnostic unit information prior 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 2.4 presents a detailed description of the variables. The numerical results are reported in Table 2.5.

Table 2.5: Why do estimates of risk aversion vary?

	Bayesian model averaging			Frequentist check (OLS)		
Variable	Post. Mean	Post. SD	PIP	Coeff.	SE	p-val.
Constant	-8.841	N.A.	1.000	-9.050	3.108	0.004
Standard error	0.980	0.035	1.000	0.980	0.070	0.000
Data characteristics						
Time span	-0.041	0.217	0.049			
Midpoint	0.003	0.084	0.014			
Panel	1.037	2.229	0.207			
Cross-section	3.424	1.866	0.833	4.098	1.841	0.026
Monthly	-0.117	0.569	0.057			
Quarterly	4.469	0.954	0.995	4.394	1.679	0.009
US	6.064	1.004	1.000	5.924	1.498	0.000
EU	0.024	0.270	0.019			
Asia	0.004	0.245	0.013			
Developing	-0.055	0.491	0.024			
Specification characteristics						
Epstein-Zin	5.488	1.370	0.991	5.592	3.390	0.099
Long-run risk	0.004	0.131	0.014			
Fixed EIS	0.024	0.276	0.020			
Nonseparable durables	4.834	1.372	0.979	5.008	3.354	0.135
Total consumption	0.207	0.801	0.080			
Exact Euler	0.063	0.345	0.045			
Human capital	0.018	0.239	0.017			
Stockholder	-5.768	1.341	0.995	-5.769	3.659	0.115
Nonstockholder	0.053	0.482	0.024			
Estimation techniques						
Experimental	-0.062	0.593	0.022			
Implied	-0.001	0.150	0.014			
GMM	-0.075	0.414	0.046			
Simulations	-0.005	0.231	0.017			
Second lag	-0.066	0.389	0.041			
Market return included	-0.116	0.486	0.070			
Consumption included	-0.195	0.628	0.108			
Publication characteristics						
Publication year	0.037	0.230	0.038			
Top journal	0.001	0.143	0.015			
Finance journal	6.358	0.949	1.000	6.297	1.565	0.000
Citations	-0.001	0.045	0.015			
Observations	1,021			1,021		
Studies	92			92		

Notes: The response variable is the reported estimate of relative risk aversion; all estimates that we collect are derived from specifications that separate risk aversion from intertemporal substitution. SD = standard deviation; PIP = posterior inclusion probability; SE = standard error. The left-hand panel applies BMA based on the unit information g-prior and the dilution model prior (Eicher et al. 2011; George 2010). See Zeugner and Feldkircher (2015) for a detailed description of the priors. The right-hand panel reports a frequentist check using ordinary least squares, which includes variables with PIPs above 0.5 in BMA. Standard errors in the frequentist check are clustered at the study level. Table 2.4 presents a detailed description of the variables.

Figure 2.5: Posterior inclusion probabilities across different prior settings



Notes: UIP = unit information prior; the prior has the same weight as one observation of data. Dilution model prior = the prior weight of each model is proportional to the determinant of the correlation matrix. BRIC and Random = the benchmark g-prior for parameters with the beta-binomial model prior for the model space, which means that each model size has equal prior probability (Fernandez et al. 2001). The HQ prior asymptotically mimics the Hannan-Quinn criterion. See Zeugner and Feldkircher (2015) for a detailed description of the priors.

the corresponding fitted values. To do so, we need to choose a specific value for each variable, which is inevitably subjective. We plug zero for the standard error to account for publication bias. To give more weight to studies with larger datasets and newer data, we plug in sample maxima for the time span and midpoint of data. We prefer if panel data, exact Euler equation, and Epstein-Zin preferences are used, first lags are not included among instruments (because of potential problems with time aggregation), the elasticity of intertemporal substitution is not fixed, and the estimate is not obtained via simulation. We also prefer if the study was published recently, in a top journal, and is frequently cited. All other variables are set to their sample means. Table 2.6 shows that such an exercise yields imprecise results, but the point estimate for economics

is still around 1, consistent with our previous results. The implied estimate for finance is somewhat larger, around 7, but not far from the 2–6 range discussed in the previous section. The implied values of risk aversion for different contexts shown in Table 2.6 lie between 1 and 7.

Table 2.6: Implied risk aversion

	Mean	95% cred. int.
Overall best practice	3.73	[-7.36, 14.82]
Economics	1.24	[-10.25, 12.73]
Finance	7.16	[-3.85, 18.17]
US	5.81	[-5.64, 17.26]
EU	1.57	[-7.07, 10.22]
Stockholder	1.49	[-6.80, 9.79]
GMM	3.79	[-6.94, 14.52]
Quarterly data	6.33	[-4.61, 17.27]

Notes: The table uses benchmark BMA results to compute relative risk aversion conditional on selected aspects of data, methodology, and publication (see text for details). That is, the table attempts to answer the question what the mean risk aversion would look like if the literature was free of publication bias and all studies used the same strategy as the one we prefer. The 95% credible intervals are reported in parentheses.

2.5 Conclusion

We provide the first meta-analysis of the literature estimating relative risk aversion. We focus on studies that use the consumption Euler equation and that break the link (present with power utility) between risk aversion and intertemporal substitution. This means that we mostly focus on estimates that employ Epstein-Zin preferences. The literature provides 1,021 estimates reported in 92 studies; we also collect 446 calibrations of relative risk aversion from 200 studies. Our results suggest a wedge between estimates and calibrations: calibrations are often larger than estimates, especially in the economics literature. The wedge increases substantially when we correct the estimates of risk aversion for publication selection bias: the corrected mean estimate is 1 for economics

and 2–7 for finance, which are the values we recommend for calibration. The finding for economics is consistent with Chetty (2006), who argues that data on labor supply behavior impose an upper bound of 2 on relative risk aversion. Our results also suggest that the estimates are systematically correlated with the context in which they are obtained, such as data dimension (time-series vs. cross-section vs. panel data), data frequency (monthly vs. quarterly vs. annual), country coverage (US vs. Europe), general form of the utility function (Epstein-Zin vs. other approaches), treatment of durables (separability vs. nonseparability), and whether or not the researcher focuses on stockholders.

Three qualifications are in order. First, our classification of studies into economics and finance fields is crude and follows the classification of journals in which the studies are published. Two studies may use a similar strategy to identify relative risk aversion, but one can be published in an economics journal, the other in a finance journal. The advantage of the journal-based classification is its clarity and parsimony; a rule based on methodology or data would also inevitably be more subjective. The sharp difference between the distribution of estimates in economics and finance according to our definition suggests that the classification we use is informative. Second, most meta-analysis methods that we use invoke the classical assumption that in the absence of publication bias there is no correlation between estimates and standard errors. The assumption does not have to hold in the risk aversion literature, because estimation approaches vary widely and some may influence both estimates and standard errors. As a partial solution we employ the p-uniform* technique, which does not need this strong assumption. The technique suggests even stronger publication bias for both economics and finance. Third, we use more than one estimate from primary studies, which violates the standard meta-analysis assumption that all estimates are independent. We partially address this problem by clustering standard errors at the study level and using wild bootstrap.

References

Albuquerque, R., Eichenbaum, M., Luo, V. X., and Rebelo, S. (2016). Valuation risk and asset pricing. *The Journal of Finance*, 71(6):2861–2904.

Ampudia, M., Cooper, R., Le Blanc, J., and Zhu, G. (2018). MPC Heterogeneity in Europe: Sources and Policy Implications. Working Paper 25082, National Bureau of Economic Research.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2018). Multiattribute utility theory, intertemporal utility, and correlation aversion. *International Economic Review*, 59(2):537–555.

Andreasen, M. M. (2012). An estimated DSGE model: Explaining variation in nominal term premia, real term premia, and inflation risk premia. *European Economic Review*, 56(8):1656–1674.

Andreasen, M. M., Fernández-Villaverde, J., and Rubio-Ramírez, J. F. (2018). The pruned state-space system for non-linear DSGE models: Theory and empirical applications. *The Review of Economic Studies*, 85(1):1–49.

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

Ascari, G., Magnusson, L. M., and Mavroeidis, S. (2021). Empirical evidence on the Euler equation for consumption in the US. *Journal of Monetary Economics*, 117(C):129–152.

Attanasio, O. P. and Weber, G. (1989). Intertemporal substitution, risk aversion and the Euler equation for consumption. *The Economic Journal*, 99(395):59–73.

Augustin, P. and Tédongap, R. (2016). Real economic shocks and sovereign credit risk. *Journal of Financial and Quantitative Analysis*, 51(2):541–587.

Bakshi, G. S. and Naka, A. (1997). An empirical investigation of asset pricing models using Japanese stock market data. *Journal of International Money and Finance*, 16(1):81–112.

Bansal, R., Gallant, A. R., and Tauchen, G. (2007a). Rational pessimism, rational exuberance, and asset pricing models. *The Review of Economic Studies*, 74(4):1005–1033.

Bansal, R., Kiku, D., and Yaron, A. (2007b). Risks for the Long Run: Estimation and Inference. Working paper, Duke University.

Bansal, R., Kiku, D., and Yaron, A. (2016). Risks for the long run: Estimation with time aggregation. *Journal of Monetary Economics*, 82(C):52–69.

Bansal, R. and Shaliastovich, I. (2013). A long-run risks explanation of predictability puzzles in bond and currency markets. *The Review of Financial Studies*, 26(1):1–33.

Bansal, R., Tallarini, T. D., and Yaron, A. (2008). The return to wealth, asset pricing and the intertemporal elasticity of substitution. Meeting papers 918, Society for Economic Dynamics.

Bansal, R. and Yaron, A. (2004). Risks for the long run: A potential resolution of asset pricing puzzles. *The journal of Finance*, 59(4):1481–1509.

Bednarek, Z. and Patel, P. (2015). Long-run Risk, Durable Consumption Growth and Estimation of Risk Aversion. Working paper, California Polytechnic State University.

Bekhtiar, K., Fessler, P., and Lindner, P. (2020). Risky assets in Europe and the US: risk vulnerability, risk aversion and economic environment. Working paper 2270, European Central Bank.

Biswas, A. and Mandal, B. (2016). Estimating Preference Parameters from Stock Returns Using Simulated Method of Moments. *Annals of Financial Economics*, 11(01):1650005.

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

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

Bretscher, L., Hsu, A., and Tamoni, A. (2020). Fiscal policy driven bond risk premia. *Journal of Financial Economics*, 138(1):53–73.

Briggs, J., Cesarini, D., Lindqvist, E., and Östling, R. (2021). Windfall gains and stock market participation. *Journal of Financial Economics*, 139(1):57–83.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and causal inference in economics. *American Economic Review*, 110(11):3634–3660.

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

Brown, A. L., Imai, T., Vieider, F., and Camerer, C. (2023). Meta-Analysis of Empirical Estimates of Loss-Aversion. *Journal of Economic Literature*, (forthcoming).

Brown, A. L. and Kim, H. (2014). Do individuals have preferences used in macro-finance models? An experimental investigation. *Management Science*, 60(4):939–958.

Bruns, S. B. and Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PloS ONE*, 11(2):e0149144.

Bufman, G. and Leiderman, L. (1990). Consumption and asset returns under non-expected utility: Some new evidence. *Economics Letters*, 34(3):231–235.

Campbell, J. Y. (1996). Understanding risk and return. *Journal of Political Economy*, 104(2):298–345.

Card, D., Kluve, J., and Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

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

Carmichael, B. and Samson, L. (1993). Excess returns determination: Empirical evidence from Canada. *Journal of Economics and Business*, 45(1):35–48.

Chen, X., Favilukis, J., and Ludvigson, S. C. (2013). An estimation of economic models with recursive preferences. *Quantitative Economics*, 4(1):39–83.

Chetty, R. (2006). A New Method of Estimating Risk Aversion. *American Economic Review*, 96(5):1821–1834.

Chiappori, P. and Paiella, M. (2011). Relative Risk Aversion Is Constant: Evidence From Panel Data. *Journal of the European Economic Association*, 9(6):1021–1052.

Cho, J. and Dokko, Y. (1993). Risk aversion in the expected and the non-expected utility functions. *Review of Quantitative Finance and Accounting*, 3(4):421–427.

Choi, H., Lugauer, S., and Mark, N. C. (2017). Precautionary saving of Chinese and US households. *Journal of Money, Credit and Banking*, 49(4):635–661.

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

Christensen, T. M. (2017). Nonparametric stochastic discount factor decomposition. *Econometrica*, 85(5):1501–1536.

Coble, K. H. and Lusk, J. L. (2010). At the nexus of risk and time preferences: An experimental investigation. *Journal of Risk and Uncertainty*, 41(1):67–79.

Colacito, R. and Croce, M. M. (2011). Risks for the long run and the real exchange rate. *Journal of Political Economy*, 119(1):153–181.

Colacito, R., Croce, M. M., Gavazzoni, F., and Ready, R. (2018). Currency risk factors in a recursive multicountry economy. *The Journal of Finance*, 73(6):2719–2756.

Constantinides, G. M. (2021). Welfare costs of idiosyncratic and aggregate consumption shocks. Working paper 29009, National Bureau of Economic Research.

Constantinides, G. M. and Ghosh, A. (2011). Asset pricing tests with long-run risks in consumption growth. *The Review of Asset Pricing Studies*, 1(1):96–136.

Cooper, R. and Zhu, G. (2016). Household finance over the life-cycle: What does education contribute? *Review of Economic Dynamics*, 20(C):63–89.

Cox, G. and Shi, X. (2023). Simple Adaptive Size-Exact Testing for Full-vector and Subvector Inference in Moment Inequality Models. *Review of Economic Studies*, 90(1):201–228.

Dave, C. and Tsang, K. P. (2014). Recursive preferences, learning and large deviations. *Economics Letters*, 124(3):329–334.

Dean, M. and Ortoleva, P. (2019). The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences*, 116(33):16262–16267.

Delikouras, S. (2017). Where's the kink? Disappointment events in consumption growth and equilibrium asset prices. *The Review of Financial Studies*, 30(8):2851–2889.

Delikouras, S. and Korniotis, G. M. (2021). Asset pricing with and without garbage: Resurrecting aggregate consumption. Working paper, University of Miami Business School.

DellaVigna, S. and Linos, E. (2022). RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *Econometrica*, 90(1):81–116.

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

Doh, T. (2013). Long-Run Risks in the Term Structure of Interest Rates: Estimation. *Journal of Applied Econometrics*, 28(3):478–497.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634.

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

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). Detecting p-hacking. *Econometrica*, 90(2):887–906.

Epstein, L. G. and Zin, S. E. (1989). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. *Econometrica*, 57(4):937–969.

Epstein, L. G. and Zin, S. E. (1991). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. *Journal of Political Economy*, 99(2):263–286.

Epstein, L. G. and Zin, S. E. (2001). The independence axiom and asset returns. *Journal of Empirical Finance*, 8(5):537–572.

Eraker, B., Shaliastovich, I., and Wang, W. (2016). Durable goods, inflation risk, and equilibrium asset prices. *The Review of Financial Studies*, 29(1):193–231.

Faria, A., Ornelas, R., and Almeida, C. (2016). Empirical selection of optimal portfolios and its influence in the estimation of Kreps-Porteus utility function parameters. *Brazilian Review of Econometrics*, 36(1):43–62.

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

Fulop, A., Heng, J., Li, J., and Liu, H. (2022). Bayesian estimation of long-run risk models using sequential Monte Carlo. *Journal of Econometrics*, 228(1):62–84.

Fulop, A., Li, J., Liu, H., and Yan, C. (2021). Estimating and Testing Long-Run Risk Models: International Evidence. Working paper, University of Manchester.

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

Gandelman, N. and Hernández-Murillo, R. (2015). Risk aversion at the country level. *Federal Reserve Bank of St. Louis Review*, 97(1):53–66.

Garcia, R. and Luger, R. (2012). Risk aversion, intertemporal substitution, and the term structure of interest rates. *Journal of Applied Econometrics*, 27(6):1013–1036.

Garcia, R., Luger, R., and Renault, E. (2003). Empirical assessment of an intertemporal option pricing model with latent variables. *Journal of Econometrics*, 116(1-2):49–83.

Garcia, R., Renault, E., and Semenov, A. (2006). Disentangling risk aversion and intertemporal substitution through a reference level. *Finance Research Letters*, 3(3):181–193.

Garcia, R., Renault, E., and Semenov, A. (2015). A Consumption CAPM with a Reference Level. Working paper, University of Montreal.

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

Ghosh, A. and Roussellet, G. (2020). Identifying beliefs from asset prices. In *Proceedings of Paris December 2019 Finance Meeting EUROFIDAI-ESSEC*. McGill University.

Gomes, F. A. R. and Ribeiro, P. F. (2015). Estimating the elasticity of intertemporal substitution taking into account the precautionary savings motive. *Journal of Macroeconomics*, 45(C):108–123.

Gomes, J. F., Kogan, L., and Yogo, M. (2009). Durability of output and expected stock returns. *Journal of Political Economy*, 117(5):941–986.

Goswami, G. and Tan, S. (2012). Pricing the US residential asset through the rent flow: A cross-sectional study. *Journal of Banking & Finance*, 36(10):2742–2756.

Goswami, G., Tan, S., and Waisman, M. (2014). Understanding the cross-section of the US housing bubble: The roles of lending, transaction costs, and rent growth. *Journal of Financial Stability*, 15(C):76–90.

Grammig, J. and Küchlin, E.-M. (2018). A two-step indirect inference approach to estimate the long-run risk asset pricing model. *Journal of Econometrics*, 205(1):6–33.

Grammig, J. and Schrimpf, A. (2009). Asset pricing with a reference level of consumption: New evidence from the cross-section of stock returns. *Review of Financial Economics*, 18(3):113–123.

Gu, L. and Huang, D. (2013). Consumption, money, intratemporal substitution, and cross-sectional asset returns. *Journal of Financial Research*, 36(1):115–146.

Guo, H. (2006). Time-varying risk premia and the cross section of stock returns. *Journal of Banking & Finance*, 30(7):2087–2107.

Hall, R. E. (1988). Intertemporal Substitution in Consumption. *Journal of Political Economy*, 96(2):339–357.

Hamori, S. (1995). On the test of the globalization of the Japanese equity market under the Kreps-Porteus preference. *Financial Engineering and the Japanese Markets*, 2(2):123–137.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hansen, L. P., Heaton, J. C., and Li, N. (2008). Consumption strikes back? Measuring long-run risk. *Journal of Political Economy*, 116(2):260–302.

Hansen, L. P. and Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286.

Hardouvelis, G. A., Kim, D., and Wizman, T. A. (1996). Asset pricing models with and without consumption data: An empirical evaluation. *Journal of Empirical Finance*, 3(3):267–301.

Hasseltoft, H. (2012). Stocks, bonds, and long-run consumption risks. *Journal of Financial and Quantitative Analysis*, 47(2):309–332.

Havranek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, 13(6):1180–1204.

Havranek, T., Horvath, R., Irsova, Z., and Rusnak, M. (2015). Cross-country heterogeneity in intertemporal substitution. *Journal of International Economics*, 96(1):100–118.

Havranek, T., Irsova, Z., Laslopova, L., and Zeynalova, O. (2023). Publication and Attenuation Biases in Measuring Skill Substitution. *Review of Economics and Statistics*, (forthcoming).

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and Van Aert, R. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, 34(3):469–475.

Horvath, R., Kaszab, L., and Marsal, A. (2021). Equity premium and monetary policy in a model with limited asset market participation. *Economic Modelling*, 95:430–440.

Huang, L., Wu, J., and Zhang, R. (2014). Exchange risk and asset returns: A theoretical and empirical study of an open economy asset pricing model. *Emerging Markets Review*, 21(C):96–116.

Hugonnier, J., Pelgrin, F., and St-Amour, P. (2013). Health and (other) asset holdings. *Review of Economic Studies*, 80(2):663–710.

Hyde, S. and Sherif, M. (2005a). Consumption asset pricing models: Evidence from the UK. *The Manchester School*, 73(3):343–363.

Hyde, S. and Sherif, M. (2005b). Don't break the habit: structural stability tests of consumption asset pricing models in the UK. *Applied Economics Letters*, 12(5):289–296.

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

Inkmann, J., Lopes, P., and Michaelides, A. (2011). How deep is the annuity market participation puzzle? *The Review of Financial Studies*, 24(1):279–319.

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

Issler, J. V. and Piqueira, N. S. (2000). Estimating relative risk aversion, the discount rate, and the intertemporal elasticity of substitution in consumption for Brazil using three types of utility function. *Brazilian Review of Econometrics*, 20(2):201–239.

Jeong, D., Kim, H., and Park, J. Y. (2015). Does ambiguity matter? Estimating asset pricing models with a multiple-priors recursive utility. *Journal of Financial Economics*, 115(2):361–382.

Jorion, P. and Giovannini, A. (1993). Time-series tests of a non-expected-utility model of asset pricing. *European Economic Review*, 37(5):1083–1100.

Kim, D. and Ryou, J. (2012). Time preference and saving rate: Implications for global imbalances. *Journal of Money and Finance*, 26(3):61–91.

Kim, H., Lee, H. I., Park, J. Y., and Yeo, H. (2010). Macroeconomic uncertainty and asset prices: A stochastic volatility model. In *AFA 2010 Atlanta Meetings Paper*. American Finance Association.

Kocherlakota, N. R. (1990). Disentangling the Coefficient of Relative Risk Aversion from the Elasticity of Intertemporal Substitution: An Irrelevance Result. *Journal of Finance*, 45(1):175–190.

Kogan, L., Papanikolaou, D., and Stoffman, N. (2020). Left behind: Creative destruction, inequality, and the stock market. *Journal of Political Economy*, 128(3):855–906.

Korniotis, G. M. (2010). Estimating panel models with internal and external habit formation. *Journal of Business & Economic Statistics*, 28(1):145–158.

Koskievic, J.-M. (1999). An intertemporal consumption–leisure model with non-expected utility. *Economics Letters*, 64(3):285–289.

Kranz, S. and Putz, P. (2022). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment. *American Economic Review*, 112(9):3124–3136.

Kuwahara, Y. and Ohkusa, Y. (1996). An alternative estimation method for the OCE model. *Applied Economics Letters*, 3(8):501–503.

Kvarven, A., Stromland, E., and Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(C):423–434.

Kwan, Y. K., Leung, C. K. Y., and Dong, J. (2015). Comparing consumption-based asset pricing models: The case of an Asian city. *Journal of Housing Economics*, 28(C):18–41.

Lee, W. (1997). Covariance risk, consumption risk, and international stock market returns. *The Quarterly Review of Economics and Finance*, 37(2):491–510.

Lence, S. H. (2000). Using consumption and asset return data to estimate farmers' time preferences and risk attitudes. *American Journal of Agricultural Economics*, 82(4):934–947.

Lybbert, T. J. and McPeak, J. (2012). Risk and intertemporal substitution: Livestock portfolios and off-take among Kenyan pastoralists. *Journal of Development Economics*, 97(2):415–426.

Maio, P. F. (2018). Does Inflation Explain Equity Risk Premia? Working paper, Hanken School of Economics.

Malloy, C. J., Moskowitz, T. J., and Vissing-Jørgensen, A. (2009). Long-run stockholder consumption risk and asset returns. *The Journal of Finance*, 64(6):2427–2479.

Mankiw, N. G. and Zeldes, S. P. (1991). The consumption of stockholders and nonstockholders. *Journal of Financial Economics*, 29(1):97–112.

McCloskey, D. N. and Ziliak, S. T. (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):356–361.

Mehra, R. and Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145–161.

Meissner, T. and Pfeiffer, P. (2022). Measuring preferences over the temporal resolution of consumption uncertainty. *Journal of Economic Theory*, 200(C):105379.

Neisser, C. (2021). The Elasticity of Taxable Income: A Meta-Regression Analysis. *Economic Journal*, 131(640):3365–3391.

Normandin, M. and St-Amour, P. (1998). Substitution, risk aversion, taste shocks and equity premia. *Journal of Applied Econometrics*, 13(3):265–281.

O'Donoghue, T. and Somerville, J. (2018). Modeling Risk Aversion in Economics. *Journal of Economic Perspectives*, 32(2):91–114.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Pommeret, A. and Epaulard, A. (2001). Agents' Preferences, the Equity Premium, and the Consumption-Saving Trade-Off: An Application to French Data. IMF Working Papers 2001/117, International Monetary Fund.

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

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

Ruge-Murcia, F. (2017). Skewness risk and bond prices. *Journal of Applied Econometrics*, 32(2):379–400.

Samson, L. and Armstrong, M. (2007). Preferences and observed risk premia: an empirical analysis. *Applied Economics Letters*, 14(6):435–439.

Schildberg-Hörisch, H. (2018). Are Risk Preferences Stable? *Journal of Economic Perspectives*, 32(2):135–154.

Schwartz, E. and Torous, W. N. (1999). Can we disentangle risk aversion from intertemporal substitution in consumption? Finance working paper 25-99, UCLA.

Semenov, A. (2003). An Empirical Assessment of a Consumption CAPM with a Reference Level under Incomplete Consumption Insurance. Working paper 2003-5, York University.

Smith, D. C. (1999). Finite sample properties of tests of the Epstein–Zin asset pricing model. *Journal of Econometrics*, 93(1):113–148.

Sönksen, J. and Grammig, J. (2021). Empirical asset pricing with multi-period disaster risk: A simulation-based approach. *Journal of Econometrics*, 222(1):805–832.

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

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):103–127.

Stanley, T. D. and Doucouliagos, H. (2010). Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys*, 24(1):170–191.

Stanley, T. D. and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78.

Stanley, T. D. and Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine*, 34(13):2116–2127.

Stanley, T. D., Doucouliagos, H., and Ioannidis, J. P. A. (2022). Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods*, 13(1):88–108.

Stanley, T. D., Doucouliagos, H., Ioannidis, J. P. A., and Carter, E. C. (2021). Detecting publication selection bias through excess statistical significance. *Research Synthesis Methods*, 12(6):776–795.

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

Stock, J. H. and Wright, J. H. (2000). GMM with weak identification. *Econometrica*, 68(5):1055–1096.

Thimme, J. and Völkert, C. (2015). Ambiguity in the cross-section of expected returns: An empirical assessment. *Journal of Business & Economic Statistics*, 33(3):418–429.

Ugur, M., Churchill, S. A., and Luong, H. M. (2020). What do we know about R&D spillovers and productivity? Meta-analysis evidence on heterogeneity and statistical power. *Research Policy*, 49(1):103866.

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

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

Van Binsbergen, J. H., Fernández-Villaverde, J., Koijen, R. S., and Rubio-Ramírez, J. (2012). The term structure of interest rates in a DSGE model with recursive preferences. *Journal of Monetary Economics*, 59(7):634–648.

Vissing-Jørgensen, A. and Attanasio, O. P. (2003). Stock-market participation, intertemporal substitution, and risk-aversion. *American Economic Review*, 93(2):383–391.

Weber, C. E. (2000). Rule-of-thumb consumption, intertemporal substitution, and risk aversion. *Journal of Business & Economic Statistics*, 18(4):497–502.

Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, 24(3):401–421.

Xu-Song, X., Li-li, M., and Ming, W. (2006). Estimation of behavior parameters based on recursive utility in asset pricing theory. In *2006 International Conference on Management Science and Engineering*, pages 282–286.

Xue, X., Reed, W. R., and Menclova, A. (2020). Social capital and health: A meta-analysis. *Journal of Health Economics*, 72(C):102317.

Yang, W. (2011). Long-run risk in durable consumption. *Journal of Financial Economics*, 102(1):45–61.

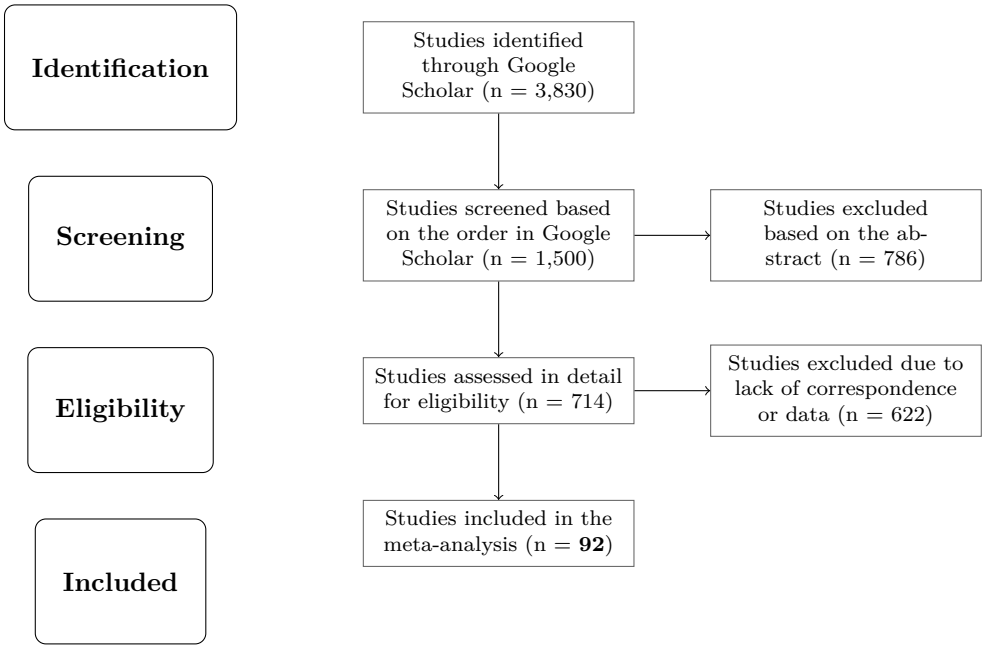
Yogo, M. (2006). A consumption-based explanation of expected stock returns. *The Journal of Finance*, 61(2):539–580.

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

Zhang, R., Brennan, T. J., and Loc, A. W. (2014). The origin of risk aversion. *Proceedings of the National Academy of Sciences*, 111(50):17777–17782.

2.A Details of Literature Search

Figure 2.A1: PRISMA flow diagram



Notes: We use the following query in Google Scholar: `"relative risk aversion" AND estimate AND ("recursive utility" OR Epstein-Zin)`. Note that Google Scholar provides fulltext search, not only the search of the title, abstract and keywords; consequently, our query is very general. For the dataset of calibrations we use the same query but replace `estimate` with `calibration`; here we inspect the studies by the order in which they are returned by Google Scholar and stop once we reach 200 usable calibration studies. The search for both estimates and calibrations was terminated on May 16, 2022. The list of the 92 studies included in the meta-analysis is available in Table 2.B1; the list of calibration studies is available in the online appendix at `meta-analysis.cz/risk`. All estimates and calibrations in our sample separate risk aversion from intertemporal substitution. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. More details on PRISMA and reporting standards of meta-analysis in general are provided by Havránek et al. (2020).

2.B Estimation of Relative Risk Aversion and Additional Summary Statistics

As we have noted, there are several ways how to estimate relative risk aversion, and a useful overview is available in Zhang et al. (2014). Potential frameworks include human subject experiments and surveys, labor-supply behavior, deductible choices in insurance contracts, auction behavior, option prices, and contestant behavior on game shows. In this paper we focus on the consumption Euler equation, which constitutes by far the most common framework used in the fields of economics and finance.

Underlying the framework is the concept of expected utility (even though, in order to separate risk aversion from intertemporal substitution, the exact form of recursive preferences used in most studies in our sample generally does not imply expected utility). The expected utility hypothesis assumes that agents in the economy are risk-averse, meaning that their preferences are concave and exhibit a diminishing marginal return utility. Hence, the degree of risk aversion is related to the curvature of the utility function. Given a form of utility function $u(c)$ where c denotes consumption, the coefficient of relative risk aversion (RRA) is defined as

$$RRA = -\frac{u''(c)}{u'(c)}c. \tag{2.B1}$$

The degree of relative risk aversion can be increasing, decreasing, or constant. In economics and finance, the largest strand of the literature employs preferences with constant relative risk aversion (CRRA), i.e., isoelastic utility (power utility function), to study agents' behavior within the economy. Measuring the structural parameters associated with household preferences, such as the coefficient of relative risk aversion and the elasticity of intertemporal substitution

(EIS), is important since they affect decisions on savings/investing and, consequently, asset prices in the economy. For instance, the degree of risk aversion plays a crucial role in the capital asset pricing model (CAPM) or consumption capital asset pricing model (CCAPM) since it heavily affects the investor's consumption and wealth portfolio, which ultimately alter asset prices.

Within the expected theory framework, a standard isoelastic utility function does not disentangle the attitude towards risk from intertemporal substitution as they are reciprocals of each other. The nonseparability of RRA and EIS ranks among the main critiques of the standard power utility function. The property means that when one of the parameters is large, the other has to be low, which is not necessarily realistic and consistent with empirical findings and commons sense. Hence, other forms of nonexpected utility must be considered to measure the degree of relative risk aversion isolated from the EIS. The most common solutions are recursive preferences of the type developed by Epstein and Zin (1989; 1991) and Weil (1989) (EZW hereinafter). This form of preferences constitutes a generalization of the standard power utility function in which the parameters governing EIS and RRA are separated. The separability of attitudes toward risk and intertemporal substitution makes the EZW recursive utility a suitable choice to estimate the degree of relative risk aversion. The EZW recursive utility function is a constant elasticity of substitution (CES) aggregator over the current and discounted future utility of consumption, taking the following form:

$$U_t = \left[(1-\beta)c_t^{1-\frac{1}{\psi}} + \beta\mu_t \left(U_{t+1}\right)^{1-\frac{1}{\psi}}\right]^{\frac{\psi}{\psi-1}}, \tag{2.B2}$$

where $0 < \beta < 1$ is the discount factor and $\psi \geq 0$ is the EIS. Households' private consumption in period t is denoted by c_t and the risk-adjusted expectation

operator is given by

$$\mu_t\left(U_{t+1}\right) = \left(\mathbb{E}_t U_{t+1}^{1-\gamma}\right)^{\frac{1}{1-\gamma}}. \tag{2.B3}$$

Employing (2.B1) with some modifications, it is straightforward to show that $\gamma \geq 0$ is the coefficient of relative risk aversion for EZW preferences. The recursive utility preferences collapse to the familiar standard CRRA utility function if $\gamma = \frac{1}{\psi}$. Additionally, when $\gamma > \frac{1}{\psi}$, the EZW preferences imply that the household prefers an early resolution of uncertainty, and a late resolution of uncertainty if $\gamma < \frac{1}{\psi}$. Assuming a representative agent model with one type of consumption goods, maximizing the intertemporal utility of the household in (2.B2) subject to an intertemporal budget constraint results in two types of Euler equations:

$$\mathbb{E}_t\left[\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\frac{1}{\psi}}\right)^{\eta}(R_{t+1}^M)^{\eta-1}R_{t+1}^i\right] = 1, \tag{2.B4}$$

and

$$\mathbb{E}_t\left[\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\frac{1}{\psi}}\right)^{\eta}(R_{t+1}^M)^{\eta}\right] = 1, \tag{2.B5}$$

where $\eta = \frac{1-\gamma}{1-\frac{1}{\psi}}$, R_{t+1}^M is the gross return on the optimal portfolio, and R_{t+1}^i is the gross return on asset i between t and $t+1$. To test the separability hypothesis, it is necessary to include the following equation

$$E_t\left[\frac{\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\frac{1}{\psi}}R_{t+1}^M\right)^{\eta}-1}{\eta}\right] = 0. \tag{2.B6}$$

Moreover, assuming that consumption growth and asset returns are jointly log-normally distributed, (2.B5) takes the form of an equivalent log-linearized version. In the log-linearized version of the equation, the riskiness of an asset depends on the conditional variance of the asset's real return, the conditional

Table 2.B1: Studies included in the meta-analysis

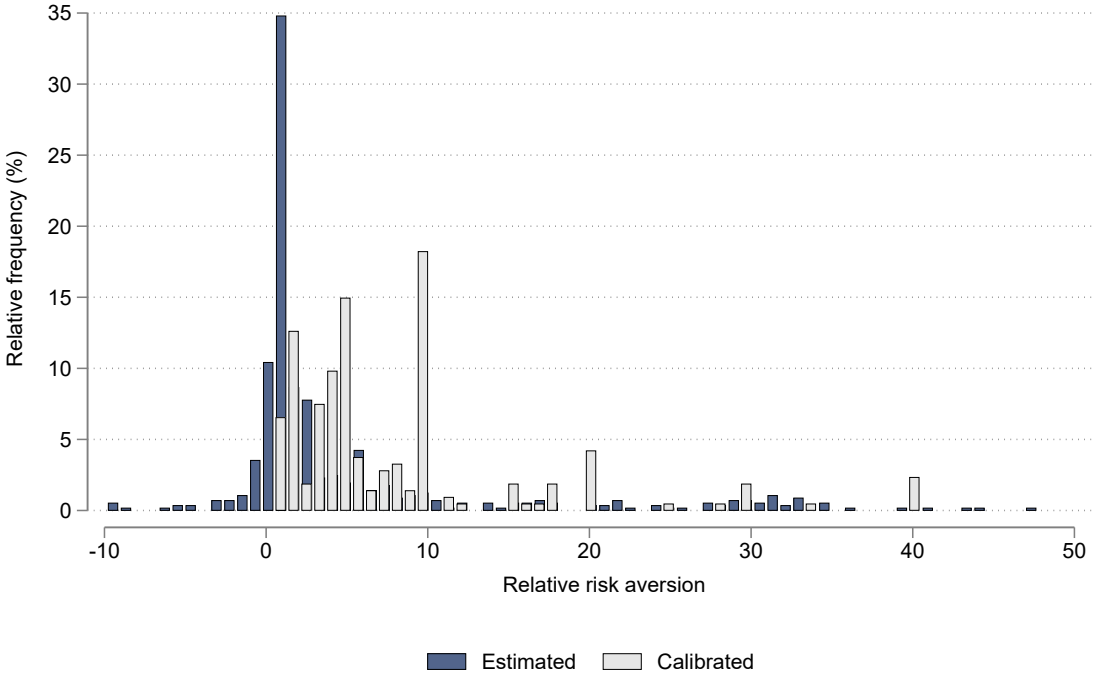
Albuquerque et al. (2016)	Dave and Tsang (2014)	Inkmann et al. (2011)
Ampudia et al. (2018)	Delikouras (2017)	Issler and Piqueira (2000)
Andersen et al. (2018)	Delikouras and Korniotis (2021)	Jeong et al. (2015)
Andreasen (2012)	Doh (2013)	Jorion and Giovannini (1993)
Andreasen et al. (2018)	Pommeret and Epaulard (2001)	Kim and Ryou (2012)
Attanasio and Weber (1989)	Epstein and Zin (1991)	Kim et al. (2010)
Augustin and Tédongap (2016)	Epstein and Zin (2001)	Kogan et al. (2020)
Bakshi and Naka (1997)	Eraker et al. (2016)	Koskievic (1999)
Bansal and Shaliastovich (2013)	Faria et al. (2016)	Kuwahara and Ohkusa (1996)
Bansal et al. (2008)	Fulop et al. (2022)	Kwan et al. (2015)
Bansal et al. (2007a)	Fulop et al. (2021)	Lee (1997)
Bansal et al. (2007b)	Garcia and Luger (2012)	Lence (2000)
Bansal et al. (2016)	Garcia et al. (2003)	Lybbert and McPeak (2012)
Bednarek and Patel (2015)	Garcia et al. (2015)	Maio (2018)
Biswas and Mandal (2016)	Ghosh and Roussellet (2020)	Malloy et al. (2009)
Bretscher et al. (2020)	Gomes and Ribeiro (2015)	Meissner and Pfeiffer (2022)
Briggs et al. (2021)	Gomes et al. (2009)	Normandin and St-Amour (1998)
Brown and Kim (2014)	Goswami and Tan (2012)	Ruge-Murcia (2017)
Bufman and Leiderman (1990)	Goswami et al. (2014)	Samson and Armstrong (2007)
Campbell (1996)	Grammig and Küchlin (2018)	Schwartz and Torous (1999)
Carmichael and Samson (1993)	Grammig and Schrimpf (2009)	Semenov (2003)
Chen et al. (2013)	Gu and Huang (2013)	Smith (1999)
Cho and Dokko (1993)	Guo (2006)	Sönksen and Grammig (2021)
Choi et al. (2017)	Hamori (1995)	Stock and Wright (2000)
Christensen (2017)	Hardouvelis et al. (1996)	Thimme and Völkert (2015)
Coble and Lusk (2010)	Hasseltoft (2012)	Van Binsbergen et al. (2012)
Colacito and Croce (2011)	Horvath et al. (2021)	Vissing-Jørgensen and Attanasio (2003)
Colacito et al. (2018)	Huang et al. (2014)	Weber (2000)
Constantinides (2021)	Hugonnier et al. (2013)	Xu-Song et al. (2006)
Constantinides and Ghosh (2011)	Hyde and Sherif (2005a)	Yogo (2006)
Cooper and Zhu (2016)	Hyde and Sherif (2005b)	

Table 2.B2: Summary statistics of benchmark calibrations

	Observations	Mean	Median	Standard deviation
All studies	200	13.13	5.93	28.62
Economics	115	16.58	5.20	36.61
Finance	85	8.47	6.00	9.14

Notes: The table only considers one benchmark calibration per study (the calibration most stressed by the authors) and only includes published studies that separate risk aversion from intertemporal substitution. Studies are classified into economics and finance categories based on the journals they were published in and using the journal classification of the Web of Science. If in the Web of Science the journal is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. Summary statistics for all calibrations from each study are reported in Table 2.1.

Figure 2.B1: Estimated and calibrated relative risk aversion in economics

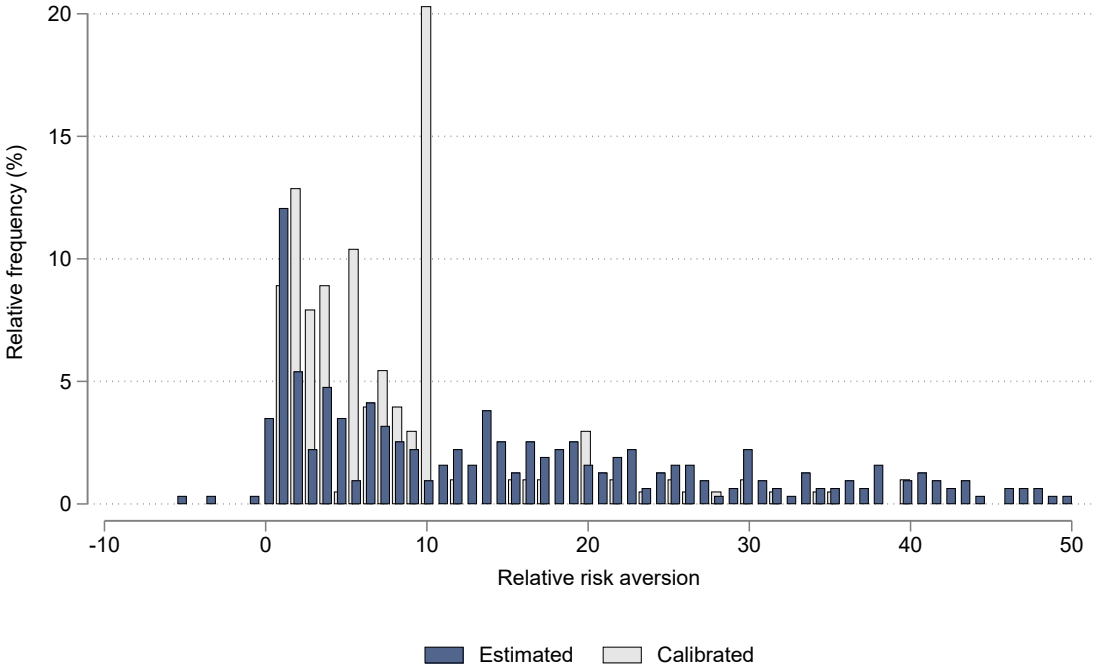


Notes: The figure shows histograms of i) 590 estimates or relative risk aversion collected from 58 economics studies and ii) 237 calibrations of relative risk aversion collected from 115 economics studies. In both cases, we only consider studies that separate risk aversion from intertemporal substitution. Studies are classified into economics and finance categories based on the journals they were published in and using the journal classification of the Web of Science. If in the Web of Science the journal is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. For ease of exposition, values below -10 and above 50 are excluded from the figure but included in all statistical tests. Summary statistics are available in Table 2.1.

covariance of the asset's real return with both consumption growth and the portfolio's real return. If the preferences reduce to the standard power utility function, i.e., $\eta = 1$, covariance risk becomes irrelevant, while in the case of EZW preferences, both covariance risk and consumption risk effectively explain assets' riskiness. Regarding the theoretical and empirical implications, Epstein and Zin (1991), Campbell (1996), and Vissing-Jørgensen and Attanasio (2003) provide more details on the log-linearized Euler equation.

The most frequently employed econometric approach to estimate the structural parameters of (2.B4) and (2.B6) or the log-linearized versions of the equa-

Figure 2.B2: Estimated and calibrated relative risk aversion in finance



Notes: The figure shows histograms of i) 431 estimates or relative risk aversion collected from 34 finance studies and ii) 209 calibrations of relative risk aversion collected from 85 finance studies. In both cases we only consider studies that separate risk aversion from intertemporal substitution. Studies are classified into economics and finance categories based on the journals they were published in and using the journal classification of the Web of Science. If in the Web of Science the journal is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. For ease of exposition, values below -10 and above 50 are excluded from the figure but included in all statistical tests. Summary statistics are available in Table 2.1.

tions is the generalized method of moments (GMM) proposed by Hansen (1982) and Hansen and Singleton (1982). Unlike other methods in the literature, the assumptions regarding the absence of heteroskedasticity and autocorrelation of residuals do not need to hold. Moreover, the GMM estimates are consistent and asymptotically efficient, unlike ordinary least squares (OLS). To implement the technique, it is necessary to identify a set of instruments that are correlated with the included endogenous variables. Market returns, stock returns, disposable income, human capital, consumption growth, and their lagged values (one-period or more) are some of the most common instruments used in the literature (see e.g., Chen et al. 2013; Faria et al. 2016; Jeong et al. 2015; Yogo

2006).

Besides OLS and GMM methods, maximum likelihood estimation (MLE) is another econometric technique used to estimate the relative risk aversion parameter (e.g., Hugonnier et al. 2013; Normandin and St-Amour 1998). Conditional on distributional assumptions, this method can provide estimates with higher statistical power than those of GMM. In the case of equilibrium models, such as dynamic stochastic general equilibrium (DSGE) models, MLE-based estimations are widely used. For instance, using an MLE procedure, Van Binsbergen et al. (2012) estimate RRA in a DSGE model with recursive preferences. The Bayesian method of estimation is another approach widely used in the literature and, in particular, DSGE models. Among others, Bretscher et al. (2020) follow a Bayesian approach to estimate the relative risk aversion parameter of EZW preferences in a New-Keynesian DSGE model. The economics literature often relies on the latter two methods to deal with investors' behavior and asset returns along with the equilibrium of the whole economy at the aggregate level. On the other hand, finance literature mainly focuses on a narrower part of the economy, i.e., the behavior of investors within the asset markets, and uses extensive data on stock market returns. Hence, the finance literature mainly employs CAPM or CCAPM models (or their extensions and alternatives) that traditionally require GMM or OLS techniques to estimate the coefficient of RRA.

Additionally, one strand of literature uses simulation-based methods to estimate the degree of risk aversion along with other structural parameters. For example, the simulated method of moments that can be considered a particular case of GMM is a widely used simulation-based technique to estimate the coefficient of relative risk aversion in the Euler equation derived from recursive preferences as it tackles the problem of aggregating consumption over time (see e.g., Albuquerque et al. 2016). Moerover, the presence of internal

habit formation in households' preferences can lead to a wedge between the RRA and the EIS as they are not the inverse of each other. Similar to models with recursive preferences, habit formation models employ estimation techniques such as GMM and OLS to estimate the coefficient of risk aversion. In this regard, Korniotis (2010) provides a detailed discussion on the estimation procedure regarding risk aversion in internal and external habit formation models. Other alternative models include expected utility with a reference level of consumption (Garcia et al. 2006), multiple-priors recursive utility with ambiguity aversion (Jeong et al. 2015), recursive preferences with smooth ambiguity aversion (Thimme and Völkert 2015), and recursive preferences with disappointment aversion (Delikouras 2017). Finally, a relatively limited literature estimates the RRA by combining the nonexpected utility model and (quasi) experimental methods. See Brown and Kim (2014) and Briggs et al. (2021) for a detailed procedure of quasi-experimental estimation of relative risk aversion in the presence of recursive preferences.

2.C Extensions and Tests of Publication Bias Models

Table 2.C1: Tests of p-hacking due to Elliott et al. (2022)

	All studies	Economics	Finance
Test for non-increasingness	0.004	0.104	1.000
Test for monotonicity and bounds	0.001	0.142	0.577
Observations ($p <= 0.15$)	755	409	346
Total observations	1,021	590	431

Notes: The table shows p-values for each test; the null hypothesis is no p-hacking, proposed by Elliott et al. (2022). The techniques rely on the conditional chi-squared test of Cox and Shi (2023). The first technique is a histogram-based test for non-increasingness of the p-curve, the second technique is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives.

Table 2.C2: Specification test for the Andrews and Kasy (2019) model

	All studies	Economics	Finance
Correlation	0.606	0.517	0.530
	[0.552, 0.656]	[0.434, 0.593]	[0.413, 0.643]

Notes: Following Kranz and Putz (2022), the table shows the correlation coefficient between the logarithm of the absolute value of the estimated inverse elasticity and the logarithm of the corresponding standard error, weighted by the inverse publication probability estimated by the Andrews and Kasy (2019) model. If the assumptions of the model hold, the correlation is zero. Bootstrapped 95% confidence interval in parentheses.

Table 2.C3: Regressing estimates on standard errors when $p < 0.005$

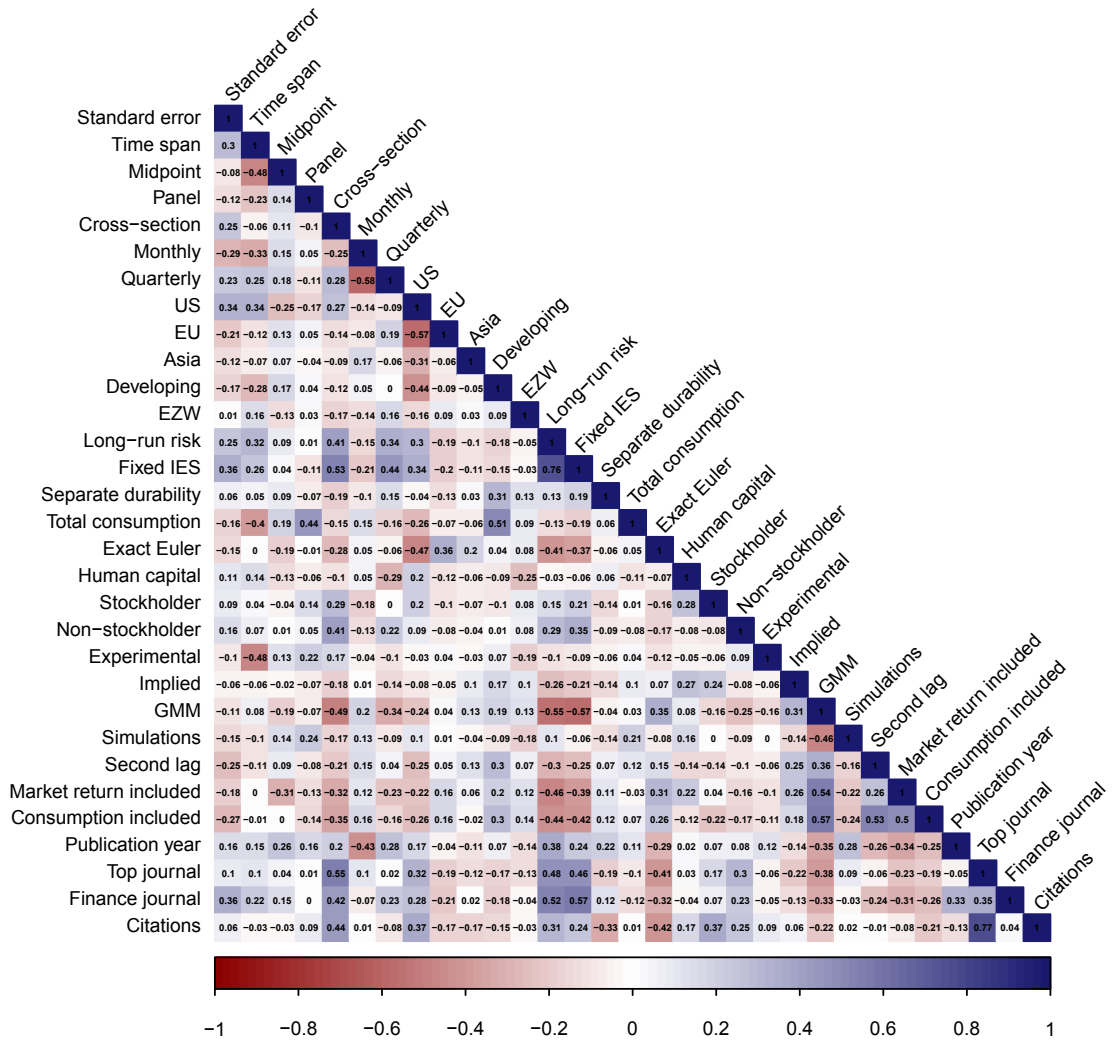
	All studies	Economics literature	Finance literature	Top journals	Implied estimate	Experimental study
Standard error	3.646***	4.171***	3.383***	3.376***	2.871***	4.261***
	(0.181)	(0.314)	(0.171)	(0.146)	(0.0134)	(0.136)
	[3.324, 4.131]	[3.016, 5.150]	[3.098, 3.866]	[2.094, 5.116]	[0.583, 4.067]	[-7.625, 16.230]
Observations	479	300	179	156	33	18
	United States	Developing country	OLS method	GMM method	Quarterly data	Annual data
Standard error	3.570***	3.722***	3.546***	3.592***	3.544***	4.033***
	(0.178)	(0.323)	(0.235)	(0.308)	(0.192)	(0.437)
	[3.259, 4.045]	[-8.862, 4.209]	[3.121, 4.325]	[2.974, 4.548]	[3.202, 4.046]	[2.848, 5.116]
Observations	327	39	155	232	247	82

Notes: The constant is included in the all regressions but not reported in the table. Standard errors, clustered at the study level, are shown in parentheses. 95% confidence intervals from the wild bootstrap are in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.D Summary Statistics, Extensions, and Additional Discussion of Heterogeneity Models

2.D.1 Variables

Figure 2.D1: Correlation matrix of BMA variables



Notes: The figure shows Pearson correlation coefficients for the variables described in Table 2.4.

Data characteristics All the variables are defined and summarized in Table 2.4 in the main body of the paper. The first category that we consider is a set of variables concerning different characteristics of the samples used

in the primary studies. We introduce eight dummy variables accounting for differences in the data. Two variables account for the difference in the data dimension: panel and cross-sectional data. Most of the reported estimates (about 76%) in primary studies are obtained using time series data, which we use as the reference category. Moreover, we codify two variables capturing data frequency. Datasets with monthly data or higher frequencies (i.e., weekly, daily) are used for 25% of the estimates, while 50% are obtained from more conventional datasets with quarterly data. Four other dummy variables denote the geographical coverage of the reported estimates. The largest group is based on the US data, accounting for 74% of estimates. The mean estimate from the US data is 31, which is substantially higher than the mean estimate of non-US data, which equals 3. This is consistent with the stream of the literature estimating higher relative risk aversion for American households compared to other countries (Gandelman and Hernández-Murillo 2015).

On the other hand, the second largest group of estimates, using European data, exhibits the opposite pattern. The European sample, comprising around 11% of the collected estimates, yields a mean around 3, while the mean estimate of non-EU datasets is 26. Two other dummy variables denote Asian and developing countries consisting of 3% and 6% of the estimates, respectively. In addition to the dummy variables, we define two variables capturing the time properties of the datasets. The first variable, time span, captures the period of data (in terms of years) used to estimate risk aversion. To control for a potential time trend reflecting structural changes in preferences (Chiappori and Paiella 2011; Schildberg-Hörisch 2018), we include the midpoint of the data as an additional explanatory variable. The earliest median year of data is 1930 in Campbell (1996), which we subtract from other studies' median years to derive a relative midpoint for each study.

Specification characteristics We codify nine dummy variables to capture different aspects of the specifications for estimating relative risk aversion. The first dummy variable denotes estimates based on the EZW recursive preferences, which are used for 90% of the estimates in our sample. The remaining 10% of the estimates are derived from other techniques that allow researchers to distinguish between risk aversion and intertemporal substitution: models with habits (Korniotis 2010), expected utility with a reference level of consumption (Garcia et al. 2006), multiple-priors recursive utility with ambiguity aversion (Jeong et al. 2015), recursive preferences with smooth ambiguity aversion (Thimme and Völkert 2015), and recursive preferences with disappointment aversion (Delikouras 2017). Next, we define a dummy variable regarding the long-run risk (LLR) model proposed by Bansal and Yaron (2004). The LLR framework contains a representative agent consumer with recursive preferences allowing for distinguishing between the RRA and EIS. The framework's other main feature is the expected consumption growth containing a small but highly persistent long-run consumption risk.

Furthermore, the LLR framework also allows for a time-varying risk premium on assets and nonindependent and identically distributed consumption growth. Using the LLR model, Hansen et al. (2008) show that the long-run risk channel can explain several problematic stylized facts in asset markets. Almost one-third (32%) of the estimates in primary studies are obtained within the LLR framework. The next variable accounts for the case when the estimated coefficients of relative risk aversion are obtained when the elasticity of intertemporal substitution is fixed in the estimation process. Around 25% of coefficients in the sample are estimated in the presence of fixed EIS. Several studies document that the estimation of EIS within a model with recursive preferences is not only empirically tricky but also irrelevant to the estimated risk aversion (e.g., Constantinides and Ghosh 2011; Malloy et al. 2009). How-

ever, there is no consensus in the literature about the exact value of the EIS, as documented by Havranek (2015) and Havranek et al. (2015).

Around 13% of the estimates are obtained in a framework where the utility function allows for nonseparability between durables and nondurables. An extensive asset pricing literature estimates the risk aversion coefficient when only nondurable goods and services are considered for consumption. There are studies, however, documenting the importance of durable goods and two-good models in estimating risk aversion (e.g., Bednarek and Patel 2015; Yang 2011). Similarly, we codify a dummy variable corresponding to the use of total consumption. Furthermore, more than one-third of the reported coefficients of RRA in our sample are estimated using a nonlinear (exact) Euler equation. The log-linearization of the Euler equation requires parametric restrictions on the structural parameters and the consumption growth and asset return, resulting in different estimates from the nonlinear case. Hence, we consider the effect of linearization of the Euler equation on the estimated risk aversion by defining a dummy variable accounting for the reported estimates obtained from the exact Euler equation.

Additionally, we add a variable to control for the role of human capital in estimating the coefficient of relative risk aversion. Since the return on human capital is not observable, it is common to use returns on equity or labor income as a proxy in the literature (Campbell 1996). Among others, Grammig and Schrimpf (2009) argue that asset pricing models augmented by human capital provide more reliable results. Slightly more than ten percent of the reported estimates are obtained using models that include human capital. Finally, two additional variables control for estimates computed exclusively for stockholders (or rich households) and nonstockholders (or poor households). Not surprisingly, as shown in Table 2.4, stockholders often show lower risk aversion than nonstockholders. The mean estimate of the coefficient of relative risk aversion

for stockholders is almost 10, while the mean estimate for nonstockholders is more than five times larger, equal to 53. Only 5% of the estimates correspond to non-stockholders and 12% for stockholders.

Estimation techniques The next category of variables considers various methods and approaches used to estimate RRA in the literature. The first dummy variable captures (quasi) experimental approaches. The variable indicates both laboratory experiments (e.g., Meissner and Pfeiffer 2022) and quasi-experimental (e.g., Lybbert and McPeak 2012) studies. The mean of such estimates is about 2, significantly lower than the mean estimate of non-experimental studies (24): though there are few (quasi) experimental studies that rely on the Euler equation. Next, we define a variable corresponding to the cases where the RRA is not directly estimated but implied by estimating other parameters in the model. The implied RRA might differ from the estimated coefficients in terms of magnitude and precision. The variable thus can be a source of heterogeneity among the estimates in the literature. The implied estimates form 12% of the sample.

Regarding the econometric approach, we define two variables capturing the techniques used in the literature. First, the GMM variable denotes the estimated coefficients obtained within the GMM framework, accounting for 59% of estimates reported in the primary studies. The second variable captures simulation-based estimates. The LLR models often employ simulation-based methods such as the simulated method of moments to estimate parameters (Hasseltoft 2012). Almost 17% of estimates in our collected sample are simulation-based. We employ the OLS estimates as the baseline category. Estimates obtained by the generalized least squares (GLS) method are also included in the baseline category. The relevance and exogeneity of instruments are essential factors affecting the reliability of estimates. We thus introduce

three dummy variables to control for the instruments used in the estimation procedure. The first variable captures estimates if the second or higher lags are included among instruments, accounting for almost 16% of estimates. We also control for the fact whether market returns are included among instruments by adding a dummy variable capturing 32% of the estimates in our sample. Finally, we include a similar dummy variable regarding the presence of consumption growth among instruments (35% of the estimates).

Publication characteristics The last group of variables reflects publication differences and measures of quality not captured by the previous variables. First, since more recent studies are more likely to provide newer methods and innovations regarding both theory and data, we control for the publication year of the estimate. Second, we categorize the estimates into economics literature and finance literature. To this end, we codify a dummy variable indicating estimates from the finance literature, which comprise 42% of the collected dataset. Studies are classified into economics and finance categories based on the journals they were published in and using the journal classification of the Web of Science. If in the Web of Science the journal is included in both categories, we follow the classification of the "most similar" journal according to the Scientific Journal Ranking. If a study is unpublished (15 studies in total), we classify it based on the prevailing publications of the corresponding author. As shown in Table 2.1, the mean of finance estimates (45) is much higher than that of our reference category, economics literature (7.5). Finally, we control for publication in top-five economics or top-three finance journals. The estimates from top journals account for 30% of the estimates reported in the primary studies. We also consider the number of citations to be a proxy for the ex-post quality of a publication and introduce a variable reflecting the number of per-year citations of each study.

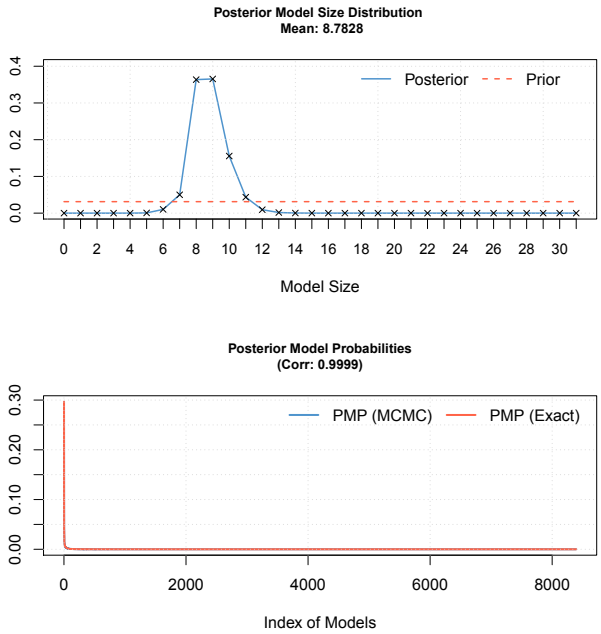
2.D.2 Results

Table 2.D1: Summary of the benchmark BMA estimation

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.794	$2 \cdot 10^6$	$1 \cdot 10^6$	2.654 mins	229,513
Modelspace	*Models visited*	*Topmodels*	*Corr PMP*	*No. Obs.*
$2.1 \cdot 10^9$	0.0011%	100	0.999	1,021
Model prior	*g-prior*	*Shrinkage-stats*		
Uniform/ 15.5	UIP	Av=0.999		

Notes: The results of this BMA specification are reported in Table 2.5. We account for collinearity among explanatory variables by employing the dilution prior suggested by George (2010); we also use the information prior recommended by Eicher et al. (2011). See Zeugner and Feldkircher (2015) for a detailed description of the priors.

Figure 2.D2: Model size and convergence for the benchmark BMA model



Notes: The figure illustrates the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table 2.5.

Table 2.D2: Results for alternative BMA priors

	BRIC g-prior			HQ g-prior		
Variable	Post. Mean	Post. SD	PIP	Post. Mean	Post. SD	PIP
Constant	-8.837	N.A.	1.000	-8.844	N.A.	1.000
Standard error	0.980	0.035	1.000	0.977	0.036	1.000
Data characteristics						
Time span	-0.044	0.225	0.052	-0.066	0.275	0.080
Midpoint	0.004	0.091	0.016	0.006	0.127	0.031
Panel	1.044	2.234	0.209	1.484	2.534	0.301
Cross-section	3.422	1.867	0.833	3.641	1.727	0.885
Monthly	-0.121	0.581	0.057	-0.200	0.724	0.099
Quarterly	4.467	0.959	0.994	4.378	0.981	0.993
US	6.068	1.007	1.000	6.108	1.052	1.000
EU	0.026	0.283	0.020	0.053	0.404	0.040
Asia	0.005	0.259	0.014	0.012	0.377	0.029
Developing	-0.055	0.495	0.025	-0.129	0.754	0.054
Specification characteristics						
Epstein-Zin	5.484	1.383	0.990	5.573	1.313	0.996
Long-run risk	0.005	0.139	0.015	0.006	0.198	0.031
Fixed EIS	0.025	0.290	0.020	0.038	0.350	0.037
Nonseparable durables	4.834	1.376	0.979	4.952	1.297	0.990
Total consumption	0.214	0.813	0.083	0.352	1.035	0.136
Exact Euler	0.068	0.360	0.049	0.142	0.518	0.096
Human capital	0.019	0.247	0.018	0.045	0.368	0.038
Stockholder	-5.767	1.350	0.995	-5.924	1.307	0.998
Nonstockholder	0.056	0.497	0.025	0.073	0.561	0.041
Estimation techniques						
Experimental	-0.071	0.640	0.025	-0.178	1.017	0.054
Implied	-0.001	0.162	0.016	0.002	0.218	0.028
GMM	-0.075	0.415	0.047	-0.093	0.461	0.066
Simulations	-0.007	0.249	0.019	-0.017	0.345	0.035
Second lag	-0.066	0.389	0.041	-0.107	0.488	0.070
Market return included	-0.119	0.492	0.072	-0.169	0.572	0.108
Consumption included	-0.199	0.633	0.110	-0.265	0.712	0.153
Publication characteristics						
Publication year	0.039	0.236	0.040	0.059	0.288	0.064
Top journal	0.001	0.151	0.016	0.001	0.217	0.033
Finance journal	6.358	0.949	1.000	6.251	0.938	1.000
Citations	-0.001	0.047	0.015	-0.002	0.068	0.031
Observations	1,021			1,021		
Studies	92			92		

Notes: The response variable is estimated relative risk aversion. SD = standard deviation, PIP = Posterior inclusion probability. The left-hand panel applies BMA based on the BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior). The right-hand panel reports the results of BMA based on HQ g-prior, which asymptotically mimics the Hannan-Quinn criterion. See Zeugner and Feldkircher (2015) for a detailed description of the priors. Table 2.4 presents a detailed description of all the variables.

Figure 2.4 in the main body of the paper illustrates the results of Bayesian model averaging. The horizontal axis denotes the cumulative posterior model probabilities, and each column corresponds to one regression model. The explanatory variables are sorted by their posterior inclusion probabilities in descending order. The blue color (darker in grayscale) and red color (lighter in grayscale) denote the positive posterior mean and negative posterior mean, respectively. A blank cell means that the variable is not included in the model. The results indicate that there are eight explanatory variables with the highest values of PIP that are likely systematically effective in explaining the size of the estimated coefficient of relative risk aversion reported in primary studies.

Table 2.5 in the main body of the paper presents the corresponding numerical results. The left panel presents BMA results for each explanatory variable by reporting posterior mean, posterior inclusion probability, and posterior standard deviation. Apart from the intercept, there are three *decisive* (according to the Raftery et al. 1997, classification) variables with PIP equal to 1 (standard error, US data, and finance journal). Four other variables have PIPs between 0.95 and 0.99 (quarterly data, stockholder, EZW preferences, and separate durability). We label these coefficients as variables with a *strong* impact. Finally, one *substantial* explanatory variable has a PIP between 0.75 and 0.95 (cross-sectional data). Additionally, Table 2.5 reports the results of the frequentist check (OLS) in the right-hand panel, including the explanatory variables with PIP larger than 0.5. The results reported in both panels are consistent since the estimated coefficients exhibit similar signs and magnitude. However, two variables estimated by OLS are marginally statistically insignificant.

Data characteristics Our findings indicate the importance of three decisive variables among data characteristics affecting the size of the estimates. First,

studies based on US data tend to report higher estimates than those of other countries. The empirical literature shows contradicting results regarding cross-country heterogeneity in risk aversion. Our BMA results are consistent with the stream of the literature indicating a higher risk aversion for the United States. Gandelman and Hernández-Murillo (2015) show that the United States has a relatively high degree of risk aversion among developed countries. On the other hand, a fraction of studies find the share of American households holding risky assets is higher than their counterparts in other countries, and this implies a lower degree of risk aversion in the United States (Bekhtiar et al. 2020).

Second, our results suggest that estimates based on cross-sectional data tend to be typically larger than the estimates obtained from time series or longitudinal data. This result is consistent with the strand of the literature concerning the cross-section of stock returns that requires a higher degree of risk aversion to reconcile aggregate consumption and market returns (see e.g., Grammig and Schrimpf 2009; Malloy et al. 2009). Significant cross-sectional variations in excess returns conflate the relationship of assets and consumption risk, which results in larger estimates of structural parameters such as the coefficient of RRA. Third, BMA results indicate that studies employing quarterly data tend to report larger estimates of relative risk aversion. On the other hand, the variable denoting frequencies higher than quarterly data, i.e., monthly frequency data, is not an insignificant explanatory variable in all BMA settings. In addition, our results suggest that the other data characteristics are not systematically correlated with the magnitude of the coefficient of relative risk aversion.

Specification characteristics Our results suggest that differences in assumed preferences may have a systematic effect on the size of the estimate. Studies that employ Epstein-Zin-Weil preferences report a higher degree of risk aver-

sion on average than those with other types of preferences, e.g., internal habit formation model. Furthermore, we find that allowing for nonseparability of durables in the utility function is associated positively with larger reported estimates. A linear combination of the discounted future nondurable and durable consumption growth determines these models' expected asset log returns. For instance, Yogo (2006) and Bednarek and Patel (2015) show that durable consumption growth plays a significant role in the pricing of stock returns, and a higher share of durable consumption in the total expenditure will result in larger estimates of relative risk aversion. Similarly, Yang (2011) finds that since both equity premium and the stock return volatility change linearly with the share of durable goods, an increase in the risk aversion coefficient can explain the increase in the premium due to the presence of durable goods in the model.

In addition, we find that stockholders are systematically less risk-averse compared to the general population. This finding aligns with the economic theory intuition that participating in stock markets indicates a lower risk aversion, while non-stockholders show a higher level of risk aversion that prevents them from holding risky assets. There is an extensive literature documenting results similar to our BMA results. Using the 17 years of data from PSID, Mankiw and Zeldes (1991) document that the implied coefficient of relative risk aversion based on stockholder consumption is one-third of those of all families in the US. Similarly, using the EZW preferences, Malloy et al. (2009) find that the risk aversion coefficient is, in general, lower for the stockholders and decreases with the level of wealth of stockholders. Their structural estimates for the stockholders and the wealthiest third of stockholders are 15 and 10, respectively. We do not find evidence that the estimates obtained within the LLR model or a nonlinear Euler equation are systematically different from the rest of the estimates. Similarly, BMA results do not show that a fixed EIS and total consumption or human capital in the estimated model systematically

affect the size of reported estimates.

Estimation techniques All variables related to the estimation approaches are negatively associated with the magnitude of reported estimates. However, the posterior mean for most of them is barely different from zero. More importantly, BMA results show that none of them is systematically important in determining the size of the coefficient of relative risk aversion. Among the variables in this category, only the variable reflecting instrumented consumption growth exhibits a PIP larger than 0.10, while the rest have PIPs between 0.01 and 0.07. These results remain the same also when we employ alternative BMA priors (Table 2.D2).

Publication characteristics Regarding the variables controlling for the quality of publications, we do not find evidence that publication year, publication in a top-five and top-three journal, or the number of citations are systematically effective in explaining the size of the reported estimates. In contrast, we confirm our previous observation that the finance literature tends to report higher estimates of RRA compared to the economics literature. BMA results indicate that finance estimates are larger than those reported in the economics literature by 6.4 on average. One explanation might be the impact of the influential studies in the finance literature. There are high-quality publications widely cited within the finance literature reporting huge estimates (e.g., Yogo 2006; Malloy et al. 2009). Such studies become benchmark studies that other researchers follow, resulting in larger estimates of the RRA coefficient.

Chapter 3

Intertemporal Substitution in Labor Supply: A Meta-Analysis

Abstract

The intertemporal substitution (Frisch) elasticity of labor supply governs how structural models predict changes in people's willingness to work in response to changes in economic conditions or government fiscal policy. We show that the mean reported estimates of the elasticity 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. A total 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 a total hours elasticity above 0.5.

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

JEL Codes: C83, E24, J21

This paper is a joint work with Tomáš Havránek, Roman Horvath, and Zuzana Irsova. The paper is published in the Review of Economic Dynamics. The authors gratefully acknowledge support from the Czech Science Foundation (project 23-05227M), the Czech Science Foundation (project 21-09231S), and the NPO Systemic Risk Institute number LX22NPO5101, funded by European Union—Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES. An online appendix with data, code, and additional results is available at `meta-analysis.cz/frisch`.

3.1 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 a key role in answering a variety of economic questions. For example, 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?

For calibrations of the elasticity in structural models, 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 and 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 total 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 total hours 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 total 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 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 and Ziliak (2019) provide an analogy to the Lombard effect in psychoacoustics: speakers involuntarily increase their effort with increasing noise. Similarly re-

[1]For recent papers on publication bias in economics, see Havranek (2015), Brodeur et al. (2016), Bruns and Ioannidis (2016), Ioannidis et al. (2017), Card et al. (2018), Christensen and Miguel (2018), Astakhov et al. (2019), DellaVigna et al. (2019), Bajzik et al. (2020), Blanco-Perez and 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 and Krueger (1995), Ashenfelter et al. (1999), and Stanley (2001).

searchers 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

[2]Recently 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.

sense only if t-statistics are symmetrically distributed. Since the t-statistic is a ratio of the point estimate to the corresponding standard error and since the symmetry property implies that the numerator and denominator are statistically independent quantities, it follows that estimates and standard errors should not be correlated. The 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 and 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 vari-

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

ation 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.

The mean elasticity is often informative 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" aggregate labor supply elasticity as a structural parameter. We control for both underlying heterogeneity (for example age, gender, and marital status) and method heterogeneity (for example time span, data frequency, and use of instrumental variables). In total we collect 23 characteristics that reflect the context in which the estimate was obtained, and we assess which variables are effective in explaining the differences in reported elasticities. For many of the method variables no established theory exists that would mandate their inclusion in the model, but anecdotal evidence still suggests they can systematically influence the reported Frisch elasticities. Hence we face substantial model uncertainty, a natural response to which in the Bayesian framework is Bayesian model averaging (see Steel 2020, for a detailed description). Given the number

of variables and need to interpret individual marginal effects, we implement Bayesian model averaging with the dilution prior suggested by George (2010), which addresses potential collinearity. As a robustness check, we use frequentist model averaging with Mallow's weights (Hansen 2007) and orthogonalize covariate space based on the approach of Amini and 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 total elasticity is the one most consistent with the literature. The total 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 meta-analysis of labor supply elasticities at the extensive mar-

gin. 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. Because of 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): intertemporal substitution is negligible at the extensive margin and small at the intensive margin. Quantitatively, though, Martinez et al. (2021) find a total hours elasticity of 0.025, while our estimate is an order of magnitude larger, about 0.25. Both numbers are very far from common calibrations of macroeconomic models. It is important to stress, however, that micro elasticities may not be fully relevant for aggregate outcomes because of aggregation and heterogeneity issues (Attanasio et al. 2018). For example, in models with heterogeneity the distribution of reservation wages matters, and it is possible to obtain large aggregate responses despite low micro elasticities.

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-from elasticities that we cover are informative regarding the real-world behavior of households with respect to temporary changes in wages.

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 the Appendix 3.A. The meta-analysis methods in both parts are identical, and any substantial differences in results are discussed in the main text. Subsection 3.4.4 in the main body of the paper summarizes the results of both meta-analyses.

3.2 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 the Appendix and conforms to the current protocols for meta-analysis in economics (Havranek et al. 2020; Irsova et al. 2024). 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 (Table 3.1; examined in the Appendix) and 762 extensive margin estimates from 38 studies (Table 3.2; examined in the main text) covering a quarter century of research on labor supply elasticities. The Aappendix also provides details on how the elasticities are estimated and how we collected estimates from individual papers.

Table 3.1: Studies included in the meta-analysis of intensive margin elasticities

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

Figure 3.1 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 impossible negative estimates sometimes appear in the literature but are very rare: a large

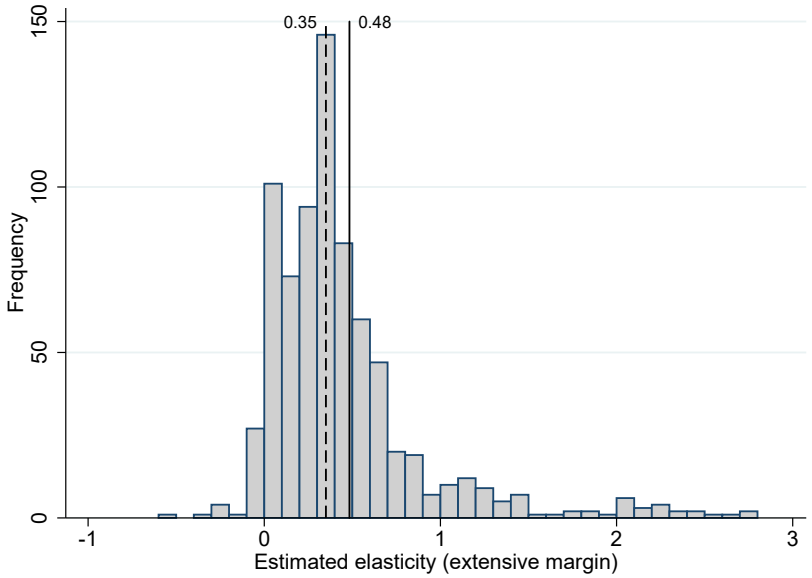
Table 3.2: Studies included in the meta-analysis of extensive margin elasticities

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

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-

Figure 3.1: Estimates are most commonly around 0.4

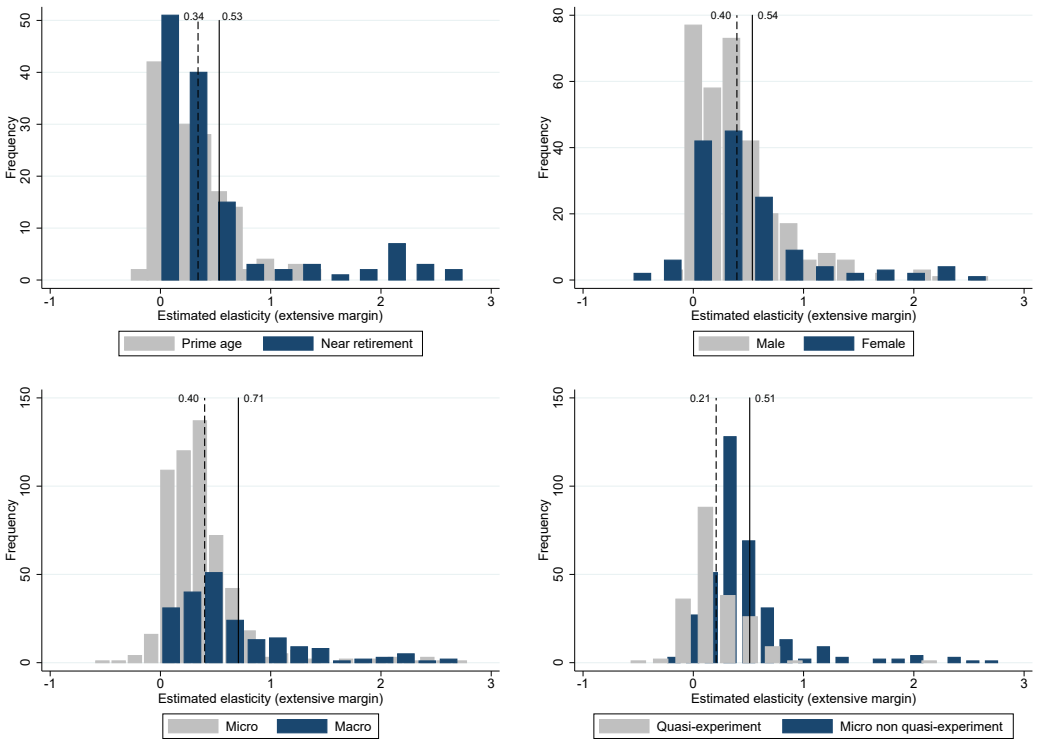


Notes: The solid line denotes the sample mean (0.48); the dashed line denotes the sample median (0.35). Estimates smaller than −1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

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 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 3.4.

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 3.2 shows

Figure 3.2: 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.

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 total hours 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 non-quasi-experimental data, which use variation in taxes or wages in the

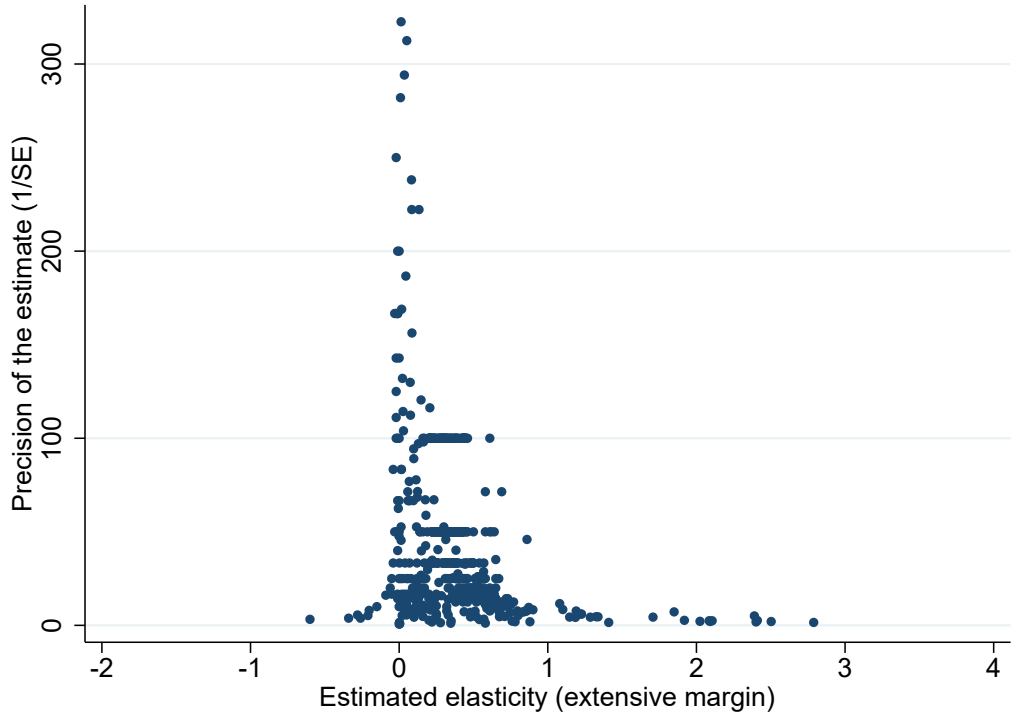
absence of significant policy shifts (0.51). These stylized facts suggest both genuine differences in the elasticity (which are however modest) and the importance of proper identification. Studies more likely to exploit truly exogenous time variation in wages are also likely to report small estimates 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.

3.3 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 3.3: 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 3.3). 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 3.3. 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:

$$\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}, \tag{3.1}$$

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 and Krueger (1995) and formalized by Stanley (2008) and Stanley and Doucouliа-

gos (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 3.3: Linear and nonlinear tests document publication bias

Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias	1.689***	0.887***	2.592***	2.173***	3.056**
(*Standard error*)	(0.264)	(0.271)	(0.530)	(0.227)	(1.500)
	[1.05, 2.36]	-	[1.55, 3.86]	[1.68, 2.70]	{0.53, 6.47}
Effect beyond bias	0.288***	0.356***	0.211***	0.243***	0.350***
(*Constant*)	(0.0442)	(0.0252)	(0.0441)	(0.0470)	(0.0463)
	[0.11, 0.37]	-	[0.06, 0.29]	[0.15, 0.34]	{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 et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)	van Aert and van Assen (2023)
Effect beyond bias	0.208***	0.354***	0.142***	0.063	0.365***
	(0.055)	(0.064)	(0.009)	(0.077)	(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. 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 3.3 presents the results of estimating Equation 3.3. 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 3.3 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.

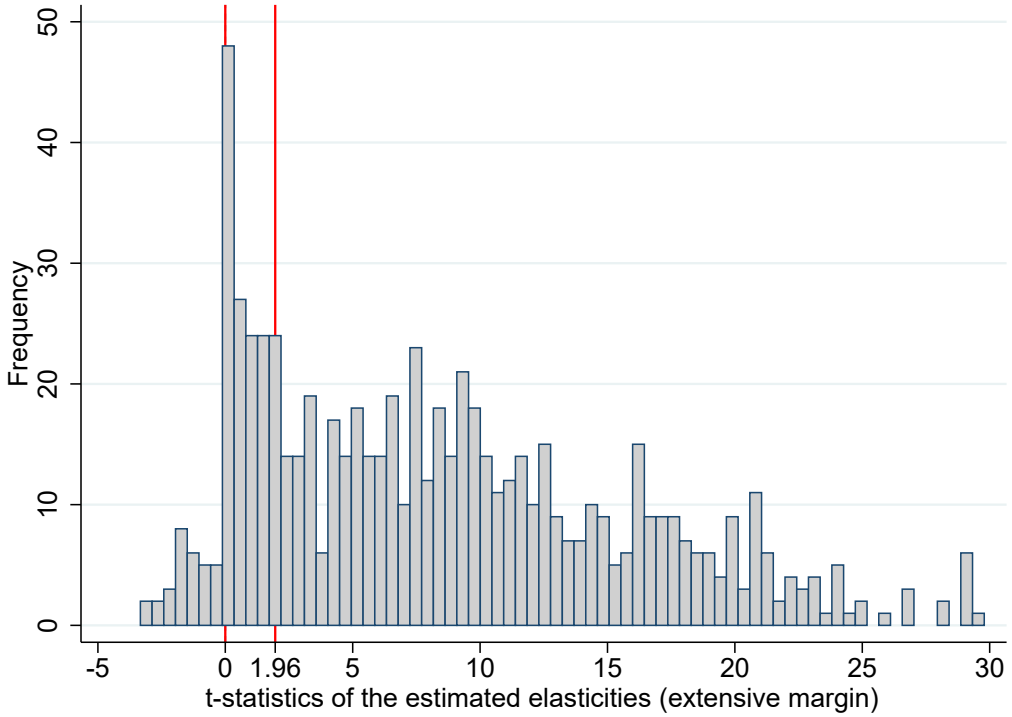
Several authors have recently addressed the nonlinearity of the funnel asymmetry test, and we use a battery of these modern techniques in panel B of

Table 3.3. 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 and 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 and 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 and 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.

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

Figure 3.4: 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.

2021) to 0.37 (van Aert and 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 extensive Frisch elasticity. We conclude that publication bias in the literature is substantial and likely to exaggerate the mean reported elasticity approximately twofold. The Appendix 3.A shows that the findings are similar for intensive margin, implying only slightly smaller publication bias. As an aside, we show in Figure 3.4 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. The Appendix 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 and 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. 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.

3.4 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 heterogene-

ity, 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.

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 and 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 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.

Table 3.4: 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
Demographic characteristics			
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
Data characteristics			
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
Specification characteristics			
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
Publication characteristics			
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.

We use the former as the baseline category and define a dummy variable for the latter.

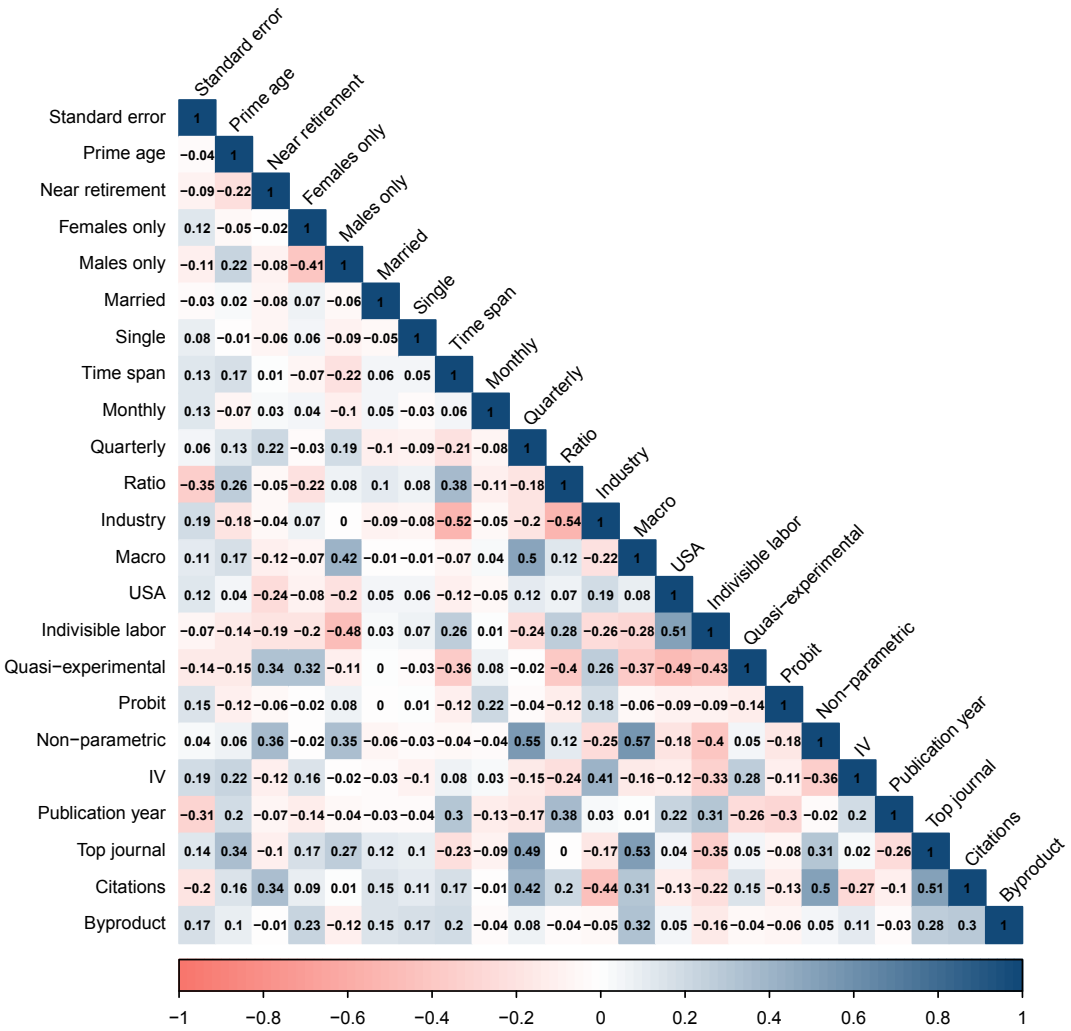
Specification characteristics We use five variables to control for the specification of primary studies. The first variable equals one if the estimate assumes the indivisibility of labor. In this case, since people can either work full-time or be unemployed, all labor fluctuations appear at the extensive 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 and Kim (2006) report the estimated Frisch extensive elasticity as a supplement.

Figure 3.5 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 3.5 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.

Figure 3.5: Correlations among explanatory variables are modest



Notes: The figure shows Pearson correlation coefficients for the variables described in Table 3.4.

3.4.2 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 and Irsova (2017); Havranek et al. (2017; 2018a;b).

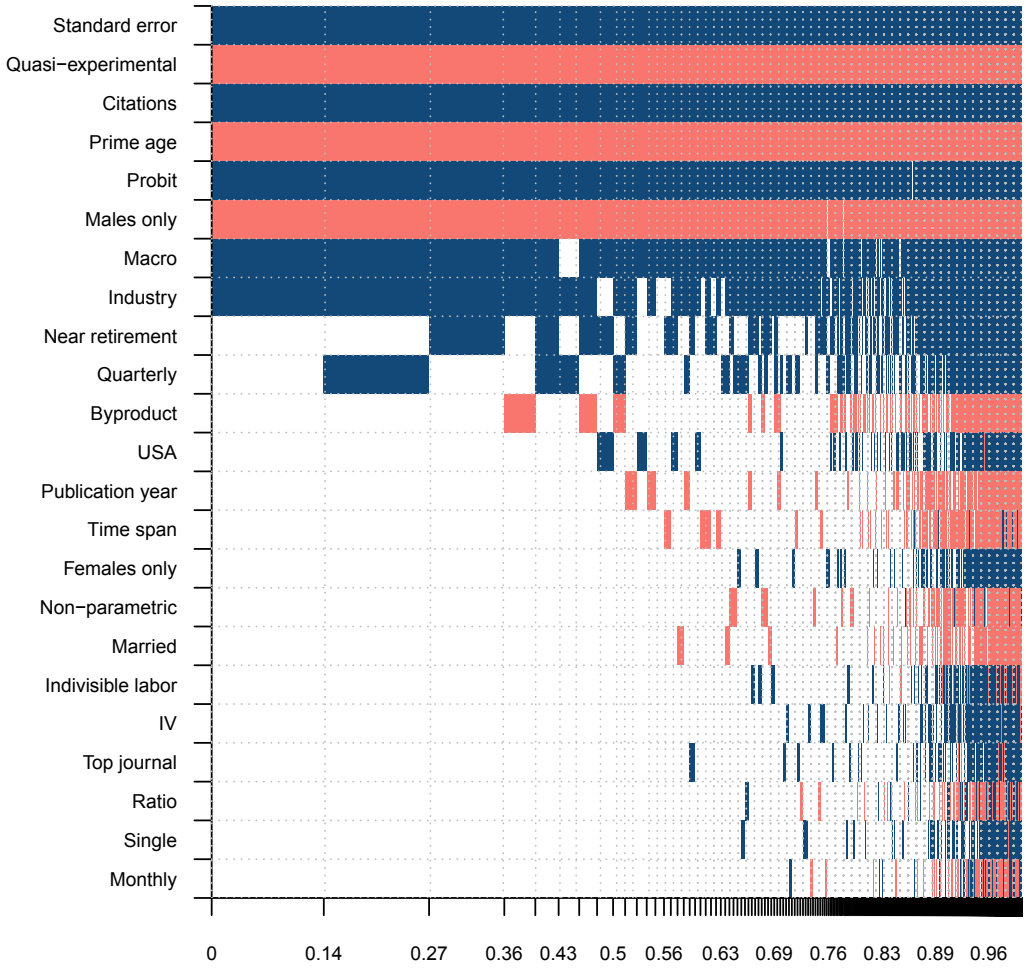
Estimating 2^{23} models would take days using a standard personal computer. Hence, we apply the Markov chain Monte Carlo algorithm (Madigan and York 1995), which goes through the models with the highest posterior model probabilities. We implement BMA using the `bms` package developed by Zeugner and 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 and 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 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.

3.4.3 Results

Figure 3.6 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 3.6 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 3.5 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 3.5 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 3.6: 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 3.4 presents a detailed description of all variables. The numerical results are reported in Table 3.5.

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

Table 3.5: Why do estimates of the elasticity vary?

Response variable: Frisch elasticity	Bayesian model averaging (baseline model)			Ordinary least squares (frequentist check)		
(extensive margin)	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
Demographic characteristics						
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			
Data characteristics						
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			
Specification characteristics						
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			
Publication characteristics						
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 3.4 presents a detailed description of all the variables.

extensive elasticities, confirming that the significant correlation between standard errors and estimates is not due to omitted aspects of demographics, data, specification, and publication.

Demographics. We find that demographic characteristics affect the estimates of the Frisch extensive elasticity in different respects. First, the estimates for men tend to be smaller than those for women. Our results also suggest that estimates of the elasticity for 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 3.1 and are also in line with the consensus in the literature. Card and Hyslop (2005), Keane (2011), and Keane and 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.

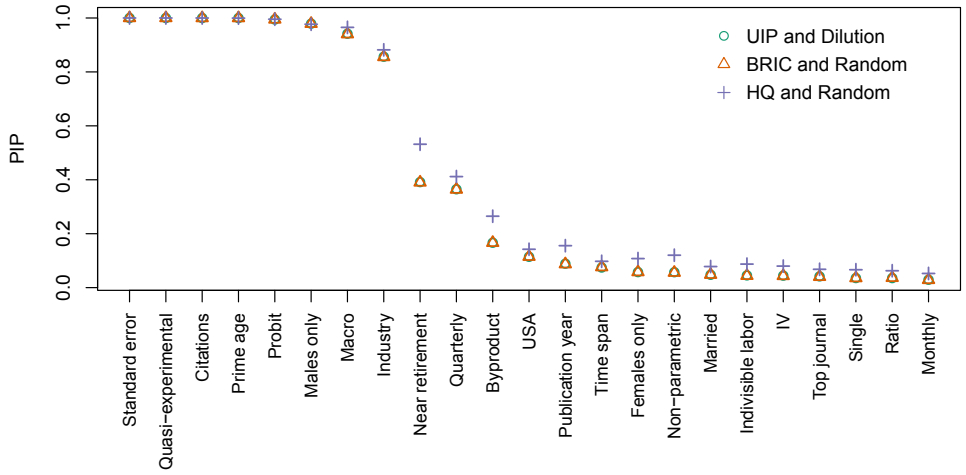
Data characteristics. Our results indicate no systematic effect of the time span, data frequency, and definition of wages used in the primary study on the reported elasticity. We do not find evidence that the US-based estimates are systematically different from estimates reported for other countries.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

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

Specifications. We find that assuming labor indivisibility is not systematically related to the size of the elasticity. The result contrasts a part of the macro literature, initiated by Hansen (1985) and Rogerson (1988), highlighting the importance of indivisible labor supply in determining the Frisch extensive elasticity. We find little evidence that either IV or non-parametric techniques used in estimating the elasticity affect the results systematically. On the other hand, elasticities estimated by the probit technique tend to be systematically larger. Finally and importantly, our results suggest that the quasi-experimental research design is a key factor for explaining the heterogeneity in the literature. Studies that do not follow the quasi-experimental approach tend to report larger estimates by 0.3 on average. This finding corroborates the pattern depicted earlier in Figure 3.1.

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 3.7: 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 and Steel 2009). We also use the BRIC g-prior suggested by Fernandez et al. (2001) together with the HQ prior. Figure 3.7 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 the appendices. Finally, we apply frequentist model averaging (FMA), which does not need priors. We use Mallow's weights (Hansen 2007) and the orthogonalization of covariate space suggested by Amini and Parmeter (2012). The robustness checks, reported in the Appendix, corroborate our main results. Regarding the analysis of heterogeneity in intensive margin elasticities examined in the Appendix 3.A, the results are similar to those for the extensive margin in several respects: publication bias is 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.

3.4.4 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 rela-

tively 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 3.6 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 total 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

Table 3.6: Mean elasticities implied by the literature

	Extensive margin		Intensive margin	
Panel A: Subjective best practice	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]
	Extensive margin		Intensive margin	
Panel B: Martinez et al. (2021)	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).

for some demographic groups: especially women and workers near retirement. For these subgroups calibrations of the total hours elasticity up to 0.4 can be backed directly 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 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.

3.5 Conclusion

A general implication of our results is that it is risky to calibrate a parameter of a structural model 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 identified by preregistered replications of estimations in economics and psychology by Open Science Collaboration (2015) and Camerer et al. (2018). So a rough 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, the results provided by Martinez et al. (2021) 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 elastic-

ities derived from direct wage measures. Although we fail to find evidence of substantial attenuation bias, we cannot rule it out.

References

Aaronson, D. and French, E. (2009). The effects of progressive taxation on labor supply when hours and wages are jointly determined. *Journal of Human Resources*, 44(2):386–408.

Altonji, J. G. (1986). Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy*, 94(3, Part 2):S176–S215.

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

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

Angrist, J. D. (1991). Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics*, 47(2-3):243–266.

Angrist, J. D., Caldwell, S., and Hall, J. V. (2021). Uber versus taxi: A driver's eye view. *American Economic Journal: Applied Economics*, 13(3):272–308.

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

Astakhov, A., Havranek, T., and Novak, J. (2019). Firm size and stock returns: A quantitative survey. *Journal of Economic Surveys*, 33(5):1463–1492.

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

Bajzik, J., Havranek, T., Irsova, Z., and Schwarz, J. (2020). Estimating the Armington elasticity: The importance of study design and publication bias. *Journal of International Economics*, 127(C).

Battisti, M., Michaels, R., and Park, C. (2023). Labor supply within the firm. *Journal of Labor Economics*, (forthcoming).

Beffy, M., Blundell, R., Bozio, A., Laroque, G., and To, M. (2019). Labour supply and taxation with restricted choices. *Journal of Econometrics*, 211(1):16–46.

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

Blanco-Perez, C. and Brodeur, A. (2020). Publication bias and editorial statement on negative findings. *The Economic Journal*, 130(629):1226–1247.

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

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

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

Borella, M., De Nardi, M., and Yang, F. (2023). Are marriage-related taxes and social security benefits holding back female labour supply? *The Review of Economic Studies*, 90(1):102–131.

Bredemeier, C., Gravert, J., and Juessen, F. (2019). Estimating labor supply elasticities with joint borrowing constraints of couples. *Journal of Labor Economics*, 37(4):1215–1265.

Brodeur, A., Carrell, S., Figlio, D., and Lusher, L. (2023). Unpacking p-hacking and publication bias. *American Economic Review*, forthcoming.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and causal inference in economics. *American Economic Review*, 110(11):3634–3660.

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

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

Bruns, S. B. and Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PloS ONE*, 11(2):e0149144.

Caldwell, S. and Oehlsen, E. (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*. PhD thesis, Massachusetts Institute of Technology.

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

N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E. J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

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

Card, D., Kluve, J., and Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

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

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

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

Chang, Y., Kim, S.-B., Kwon, K., and Rogerson, R. (2011). Interpreting labor supply regressions in a model of full-and part-time work. *American Economic Review*, 101(3):476–481.

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

Chetty, R. (2012). Bounds on Elasticities With Optimization Frictions: A

Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica*, 80(3):969–1018.

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

Chetty, R., Looney, W., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–77.

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

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

Domeij, D. and Floden, M. (2006). The labor-supply elasticity and borrowing constraints: Why estimates are biased. *Review of Economic Dynamics*, 9(2):242–262.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634.

Ehrenbergerova, D., Bajzik, J., and Havranek, T. (2023). When Does Monetary Policy Sway House Prices? A Meta-Analysis. *IMF Economic Review*, 71(2):538–573.

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

Elliott, G., Kudrin, N., and Wuthrich, K. (2022). Detecting p-hacking. *Econometrica*, 90(2):887–906.

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

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

Fabo, B., Jancokova, M., Kempf, E., and Pastor, L. (2021). Fifty shades of QE: Comparing findings of central bankers and academics. *Journal of Monetary Economics*, 120(C):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):1975–2026.

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

Fiorito, R. and Zanella, G. (2012). The anatomy of the aggregate labor supply elasticity. *Review of Economic Dynamics*, 15(2):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):395–427.

French, S. and Stafford, T. (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.

Gechert, S., Havranek, T., Irsova, Z., and Kolcunova, D. (2022). Measuring Capital-Labor Substitution: The Importance of Method Choices and Publication Bias. *Review of Economic Dynamics*, 45:55–82.

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

Gerber, A., Malhotra, N., 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):313–326.

Gine, X., Martinez-Bravo, M., and Vidal-Fernandez, M. (2017). Are labor supply decisions consistent with neoclassical preferences? Evidence from Indian boat owners. *Journal of Economic Behavior & Organization*, 142(C):331–347.

Gourio, F. and Noual, P.-A. (2009). The marginal worker and the aggregate elasticity of labor supply. Working Papers Series 2006-009, Boston University Dept. of Economics.

Gruber, J. and Wise, D. A. (1999). *Social security and retirement around the world*. University of Chicago Press.

Haan, P. and Uhlendorff, A. (2013). Intertemporal labor supply and involuntary unemployment. *Empirical Economics*, 44(2):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):281–323.

Ham, J. C. and Reilly, K. T. (2002). Testing intertemporal substitution, im-

plicit contracts, and hours restriction models of the labor market using micro data. *American Economic Review*, 92(4):905–927.

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

Hansen, G. D. (1985). Indivisible labor and the business cycle. *Journal of Monetary Economics*, 16(3):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):57–67.

Havranek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, 13(6):1180–1204.

Havranek, T., Herman, D., and Irsova, Z. (2018a). Does daylight saving save electricity? A meta-analysis. *The Energy Journal*, 39(2):35–62.

Havranek, T. and Irsova, Z. (2017). Do borders really slash trade? A meta-analysis. *IMF Economic Review*, 65(2):365–396.

Havranek, T., Irsova, Z., Laslopova, L., and Zeynalova, O. (2023). Publication and attenuation biases in measuring skill substitution. *The Review of Economics and Statistics*, (forthcoming).

Havranek, T., Irsova, Z., and Vlach, T. (2018b). Measuring the income elasticity of water demand: the importance of publication and endogeneity biases. *Land Economics*, 94(2):259–283.

Havranek, T., Rusnak, M., and Sokolova, A. (2017). Habit formation in consumption: A meta-analysis. *European Economic Review*, 95(C):142–167.

Havranek, T., Stanley, T. D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and van Aert, R. C. M. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, 34(3):469–475.

Imai, T., Rutter, T. A., and Camerer, C. F. (2021). Meta-analysis of present-bias estimation using convex time budgets. *The Economic Journal*, 131(636):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., Stanley, T. D., and Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, 127(605):F236–F265.

Irsova, Z., Bom, P. R. D., Havranek, T., and Rachinger, H. (2023). Spurious Precision in Meta-Analysis. CEPR Discussion Papers 17927, Centre for Economic Policy Research.

Irsova, Z., Doucouliagos, H., Havranek, T., and Stanley, T. (2024). Meta-Analysis of Social Science Research: A Practitioner's Guide. *Journal of Economic Surveys*, (forthcoming).

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

Keane, M. and Neal, T. (2023). Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics*, 235(2):1625–1653.

Keane, M. and Rogerson, R. (2015). Reconciling micro and macro labor supply

elasticities: A structural perspective. *Annual Review of Economics*, 7(1):89–117.

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

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

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

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

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

Lee, C.-I. (2001). Finite sample bias in IV estimation of intertemporal labor supply models: Is the intertemporal substitution elasticity really small? *The Review of Economics and Statistics*, 83(4):638–646.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674.

Looney, A. and Singhal, M. (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):1059–1085.

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

Manoli, D. and Weber, A. (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. and Weber, A. (2016). Nonparametric evidence on the effects of financial incentives on retirement decisions. *American Economic Journal: Economic Policy*, 8(4):160–182.

Martinez, I. Z., Saez, E., and Siegenthaler, M. (2021). Intertemporal labor supply substitution? Evidence from the Swiss income tax holidays. *American Economic Review*, 111(2):506–546.

Matousek, J., Havranek, T., and Irsova, Z. (2022). Individual discount rates: a meta-analysis of experimental evidence. *Experimental Economics*, 25(1):318–358.

McCloskey, D. N. and Ziliak, S. T. (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):356–361.

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):619–634.

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

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):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):aac4716.

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

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

Pistaferri, L. (2003). Anticipated and unanticipated wage changes, wage risk, and intertemporal labor supply. *Journal of Labor Economics*, 21(3):729–754.

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

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

Rusnak, M., Havranek, T., and Horvath, R. (2013). How to solve the price puzzle? A meta-analysis. *Journal of Money, Credit and Banking*, 45(1):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):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):683–710.

Stanley, T. and Doucouliagos, H. (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):131–150.

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys*, 19(3):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):103–127.

Stanley, T. D., Doucouliagos, H., and Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36(10):1580–1598.

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

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58(3):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):103765.

van Aert, R. C. and van Assen, M. (2023). Correcting for publication bias in a meta-analysis with the p-uniform* method. Working paper, Tilburg University & Utrecht University.

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):577–591.

Whalen, C. and Reichling, F. (2017). Estimates of the Frisch elasticity of labor supply: A review. *Eastern Economic Journal*, 43(1):37–42.

Yang, F., Havranek, T., Irsova, Z., and Novak, J. (2023). Is Research on Hedge Fund Performance Published Selectively? A Quantitative Survey,. *Journal of Economic Surveys*, (forthcoming).

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

Zigraiova, D., Havranek, T., Irsova, Z., and Novak, J. (2021). How puzzling is the forward premium puzzle? A meta-analysis. *European Economic Review*, 134(C):103714.

Ziliak, J. P. and Kniesner, T. J. (2005). The effect of income taxation on consumption and labor supply. *Journal of Labor Economics*, 23(4):769–796.

3.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; 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 3.B2. We find 40 studies, listed in Table 3.A1, which together provide 709 estimates of the intensive margin elasticity; details on the extraction of estimates from individual studies are available in Table 3.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 3.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 3.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 3.A3 and Table 3.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 3.A3, we repeat the analysis of publication bias previously reported in Table 3.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 and Neal (2023) show that for instrumental variables, estimates and standard er-

rors 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 3.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 3.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 3.A4. Table 3.A5 and Figure 3.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 3.A7.

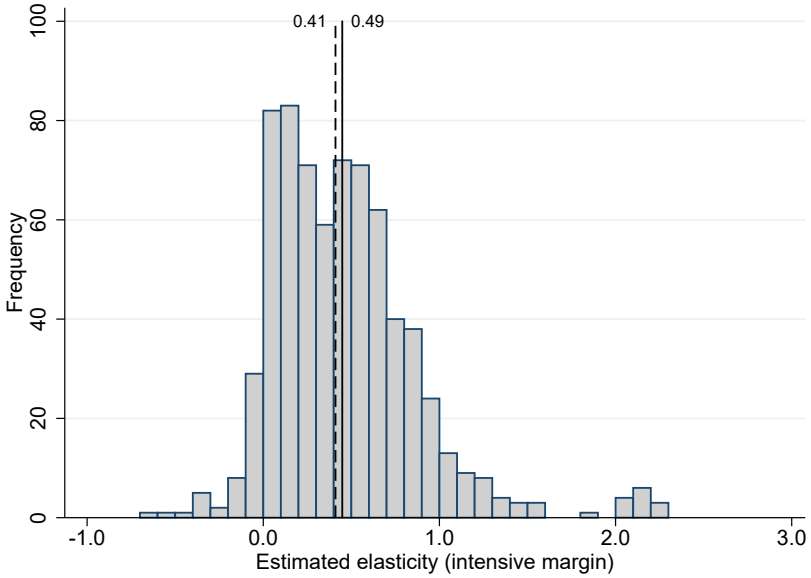
The bottom line of the meta-analysis of intensive margin elasticities is reported, together with the corresponding evidence for the extensive margin, in the main body of the paper. The table presents implied elasticities in various contexts: that is, mean elasticities corrected for publication bias and 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 3.A1: Studies included in the meta-analysis of intensive margin elasticities

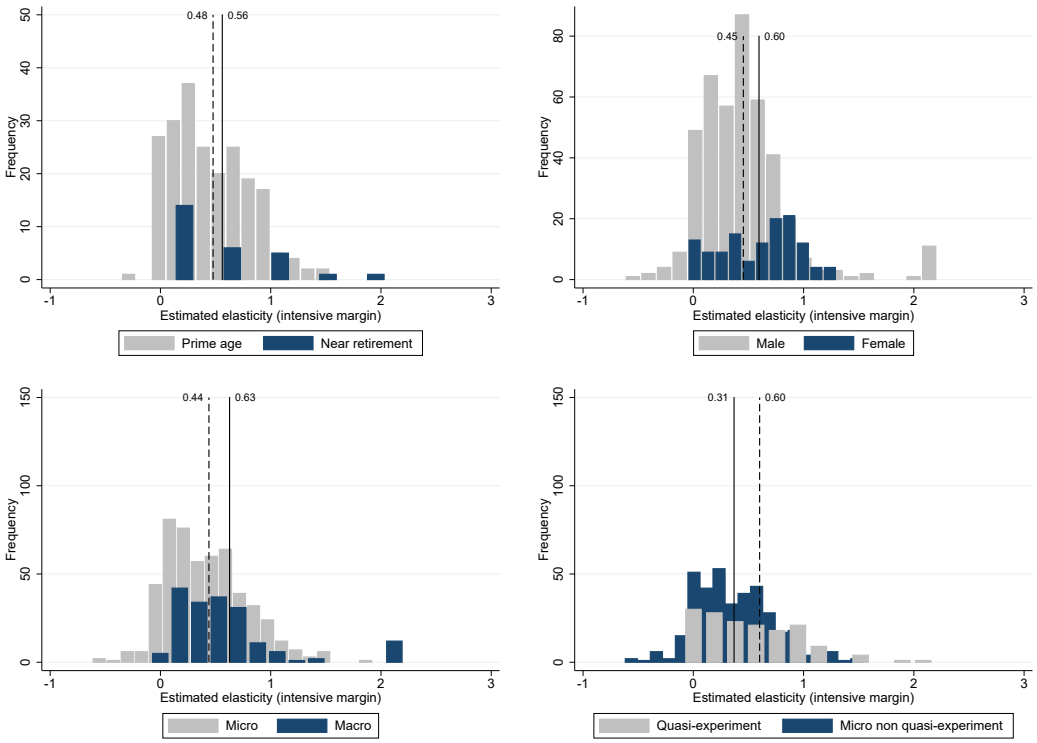
Aaronson and French (2009)	Ham and Reilly (2002)
Altonji (1986)	Inoue (2015)
Angrist (1991)	Karabarbounis (2016)
Angrist et al. (2021)	Keane and Wasi (2016)
Attanasio et al. (2018)	Kimmel and Kniesner (1998)
Battisti et al. (2023)	Kneip et al. (2019)
Beffy et al. (2019)	Kuroda and Yamamoto (2008)
Blundell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

Figure 3.A1: Estimates between 0 and 0.7 are almost equally common



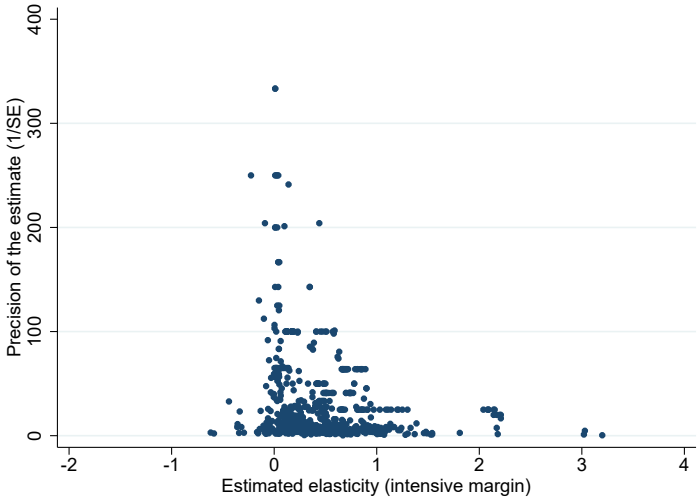
Notes: The solid line denotes the sample mean (0.49); the dashed line denotes the sample median (0.41). Note the jump at 0. Estimates smaller than -1 and larger than 3 are excluded from the figure for ease of exposition but included in all tests.

Figure 3.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 3.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 3.A2: Linear and nonlinear tests document publication bias

Panel A: Linear tests					
	OLS	FE	Precision	Study	MAIVE
Publication bias	0.590**	0.928***	1.179***	0.780***	5.163**
(*Standard error*)	(0.266)	(0.110)	(0.440)	(0.257)	(2.159)
	[-0.01, 1.22]	-	[0.23, 2.17]	[0.23, 1.38]	{0.73, 3.72}
Effect beyond bias	0.373***	0.329***	0.297***	0.331***	0.279***
(*Constant*)	(0.0567)	(0.0170)	(0.0666)	(0.0467)	(0.0505)
	[0.24, 0.49]	-	[0.14, 0.50]	[0.23, 0.43]	{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 et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)	van Aert and van Assen (2023)
Effect beyond bias	0.199***	0.295***	0.213***	0.343***	0.387***
	(0.045)	(0.003)	(0.014)	(0.126)	(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 (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 intervals. Panel B presents the mean elasticity corrected for publication bias using nonlinear techniques. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.A3: Publication bias in subsamples of the literature

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

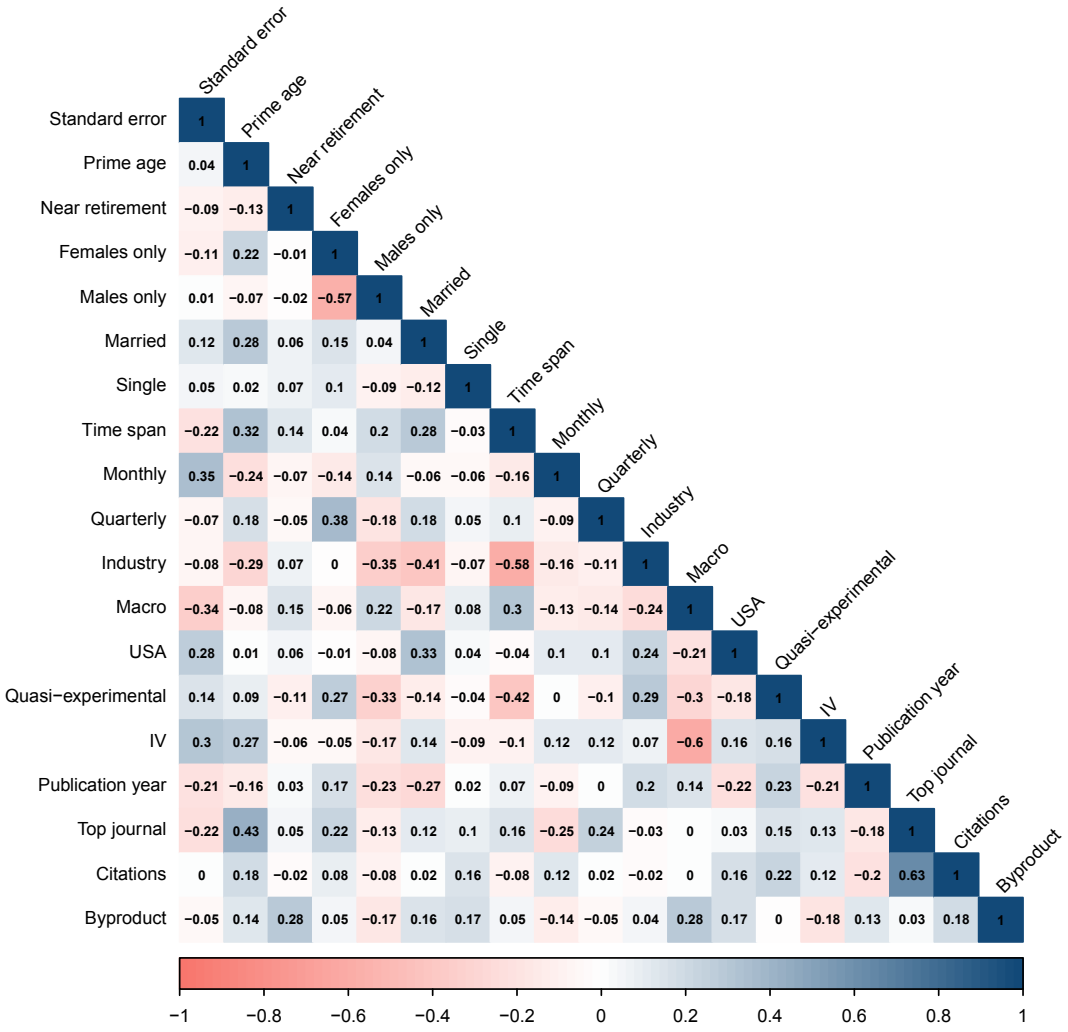
Notes: Panel A presents the results of regression $\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}$, where $\hat{\eta}_{ij}$ and $SE(\hat{\eta}_{ij})$ are the i-th estimated Frisch intensive margin elasticity and its standard error reported in the j-th study. OLS = ordinary least 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 3.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
Demographic characteristics			
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
Data characteristics			
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
Specification characteristics			
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
Publication characteristics			
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 3.A4: Correlations among explanatory variables



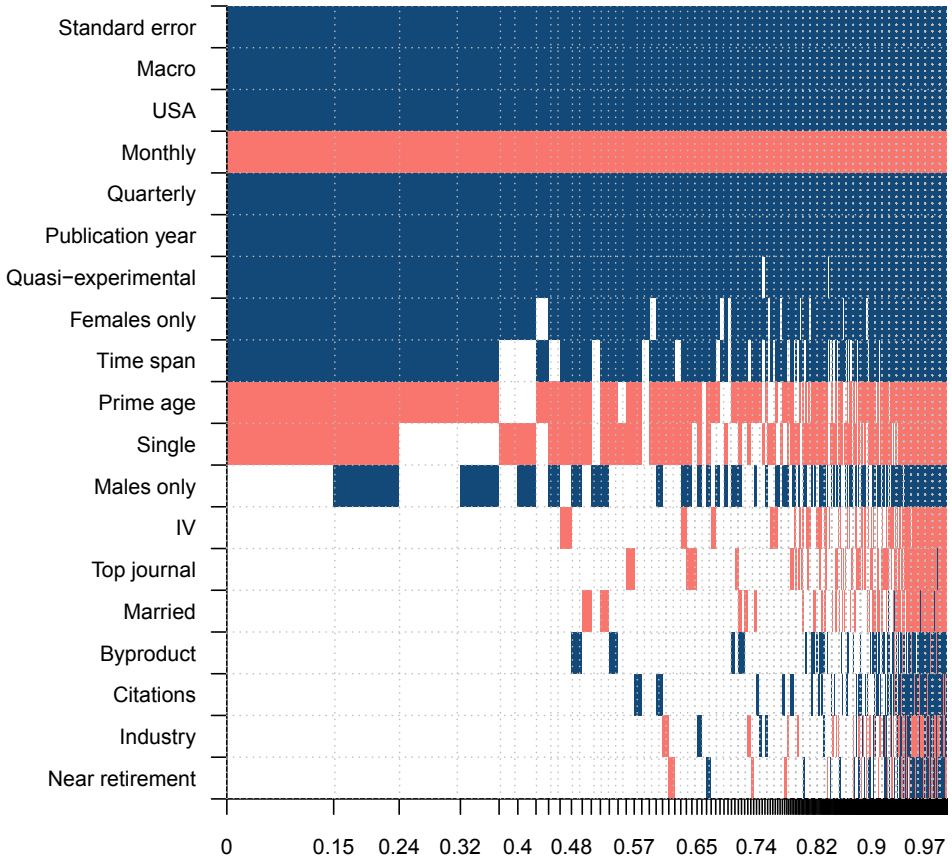
Notes: The figure shows Pearson correlation coefficients for the variables described in Table 3.A4; only intensive margin elasticities are used for the computation.

Table 3.A5: Why do estimates of the elasticity vary?

Response variable: Frisch elasticity	Bayesian model averaging (baseline model)			Ordinary least squares (frequentist check)		
(intensive margin)	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
Demographic characteristics						
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			
Data characteristics						
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
Specification characteristics						
Quasi-experimental	0.157	0.045	0.988	0.171	0.080	0.039
IV	-0.003	0.014	0.101			
Publication characteristics						
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		

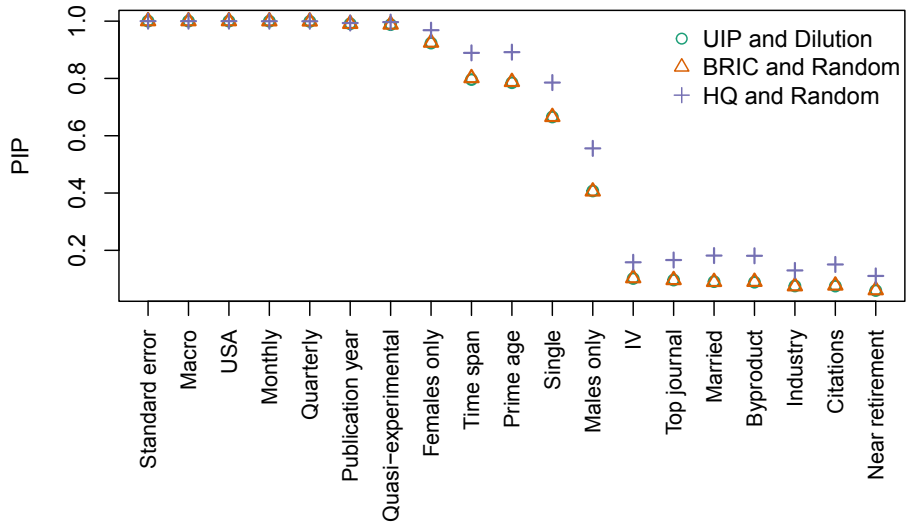
Notes: The response variable is the Frisch elasticity of labor supply at the intensive margin. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = Posterior inclusion probability, SE = standard error. The left-hand panel applies BMA based on the UIP g-prior and the dilution prior (Eicher 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 3.A4 presents a detailed description of all the variables.

Figure 3.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 3.A4 presents a detailed description of all variables. The numerical results are reported in Table 3.A7.

Figure 3.A6: Posterior inclusion probabilities hold across different priors



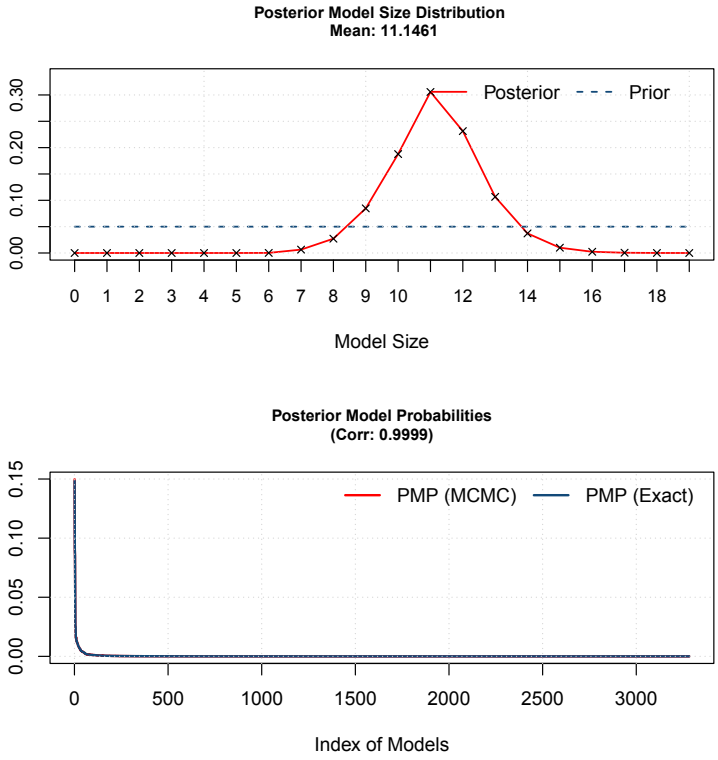
Notes: UIP and Dilution = priors according to Eicher et al. (2011) and George (2010). BRIC and Random = the benchmark g-prior for parameters with the beta-binomial model prior (each model size has equal prior probability). The HQ prior asymptotically mimics the Hannan-Quinn criterion. PIP = posterior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

Table 3.A6: Summary of the BMA estimation (UIP and dilution prior)

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

Notes: The results of this BMA specification are reported in Table 3.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 3.A7: Model size and convergence in the BMA model (UIP and dilution prior)



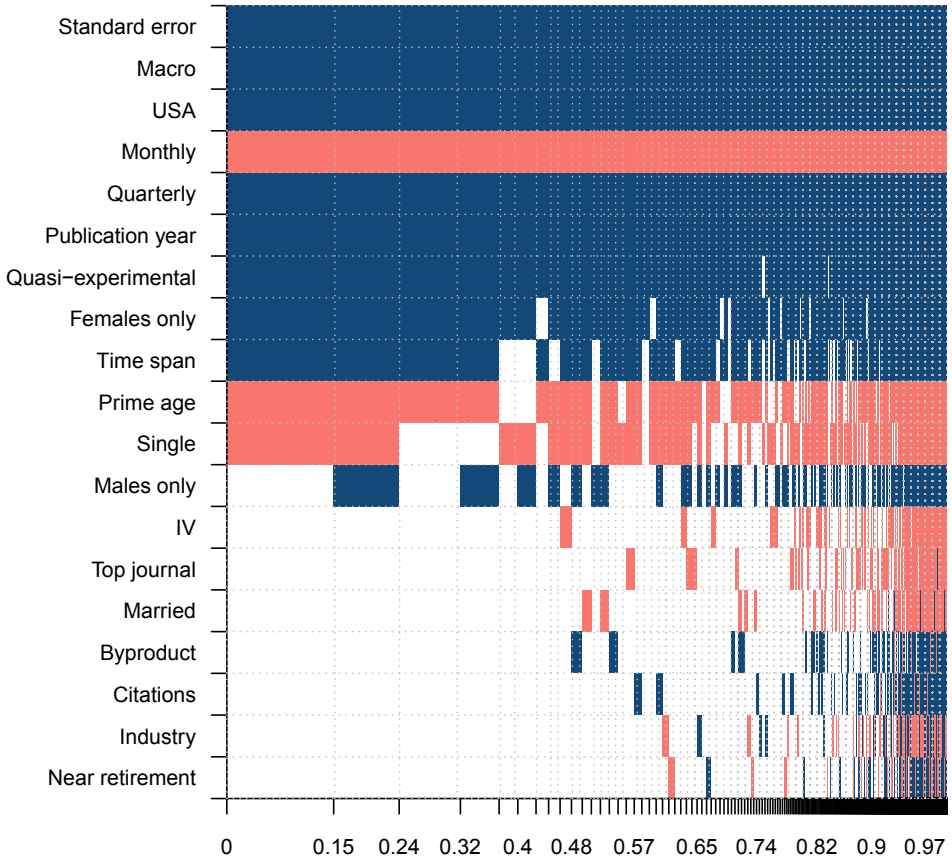
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in Table 3.A5.

Table 3.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
Demographic characteristics									
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
Data characteristics									
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
Specification characteristics									
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
Publication characteristics									
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 3.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 and Parmeter (2012) to conduct the frequentist model averaging exercise.

Figure 3.A8: Model inclusion in Bayesian model averaging (Random and BRIC)



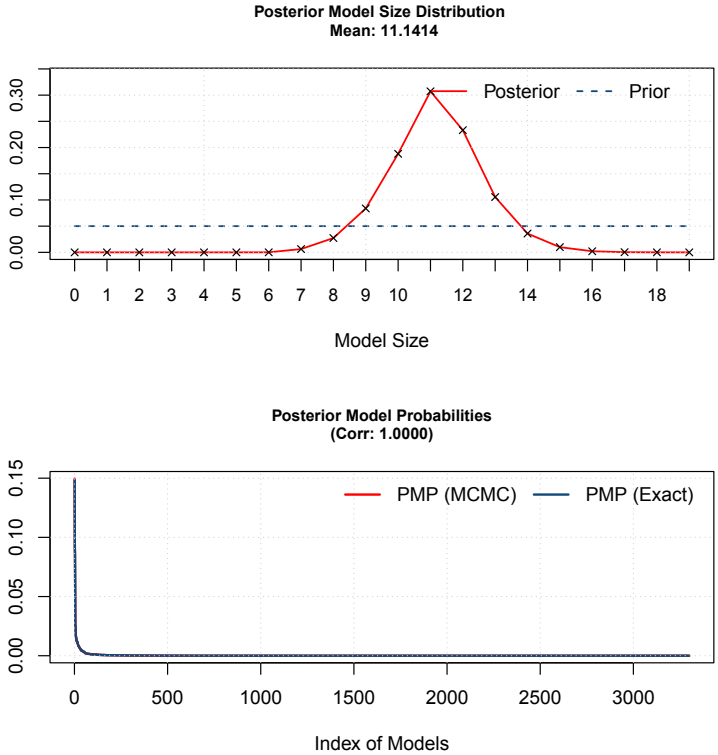
Notes: The response variable is the estimate of the Frisch elasticity of labor supply at the intensive margin. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 3.A7.

Table 3.A8: Summary of the BMA (Random and BRIC)

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

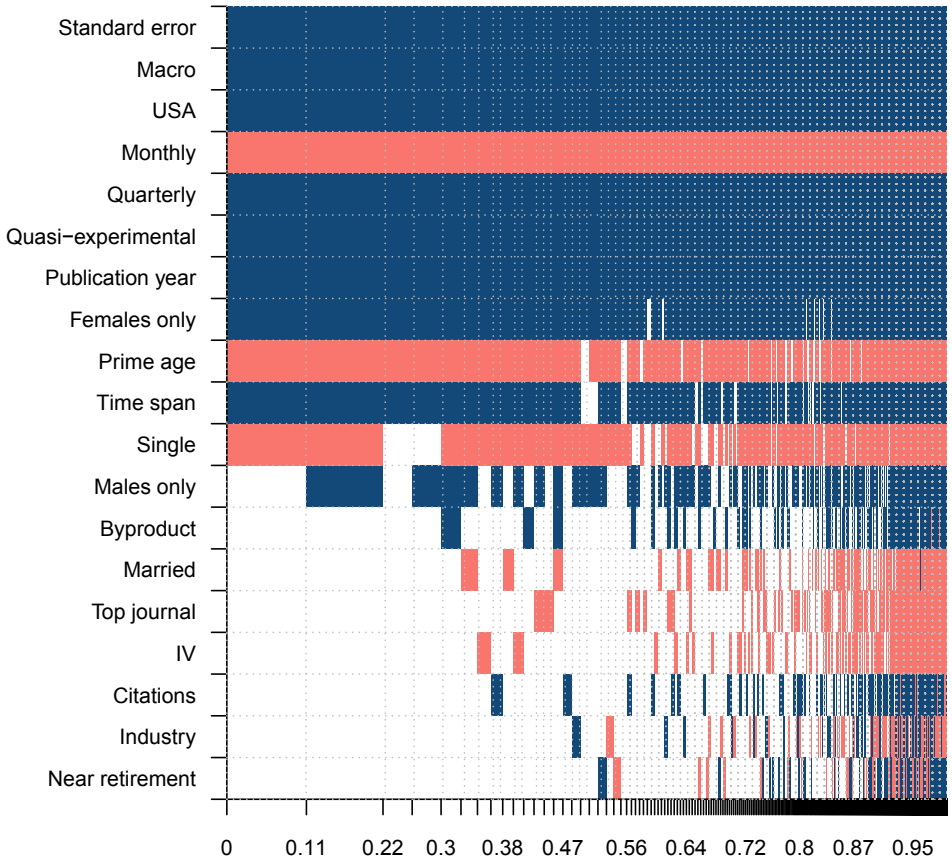
Notes: The results of this BMA specification are reported in Table 3.A7. The estimation is based on BRIC g-prior suggested by Fernandez et al. (2001) and the beta-binomial model prior according to Ley and Steel (2009).

Figure 3.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 3.A7.

Figure 3.A10: Model inclusion in BMA (Random and HQ g-prior)



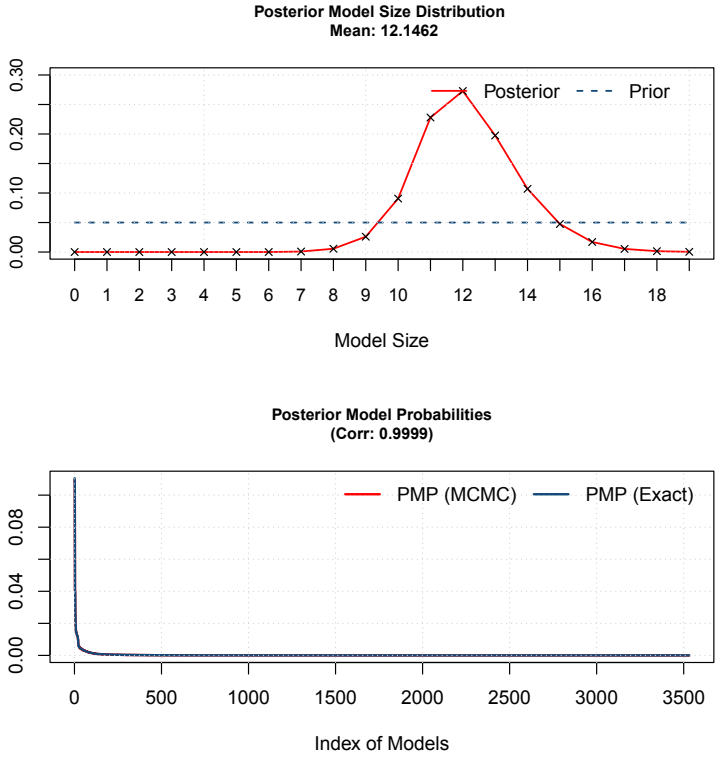
Notes: The response variable is the estimate of the Frisch intensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 3.A7.

Table 3.A9: Summary of the BMA (Random and HQ g-prior)

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

Notes: The results of this BMA specification are reported in Table 3.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).

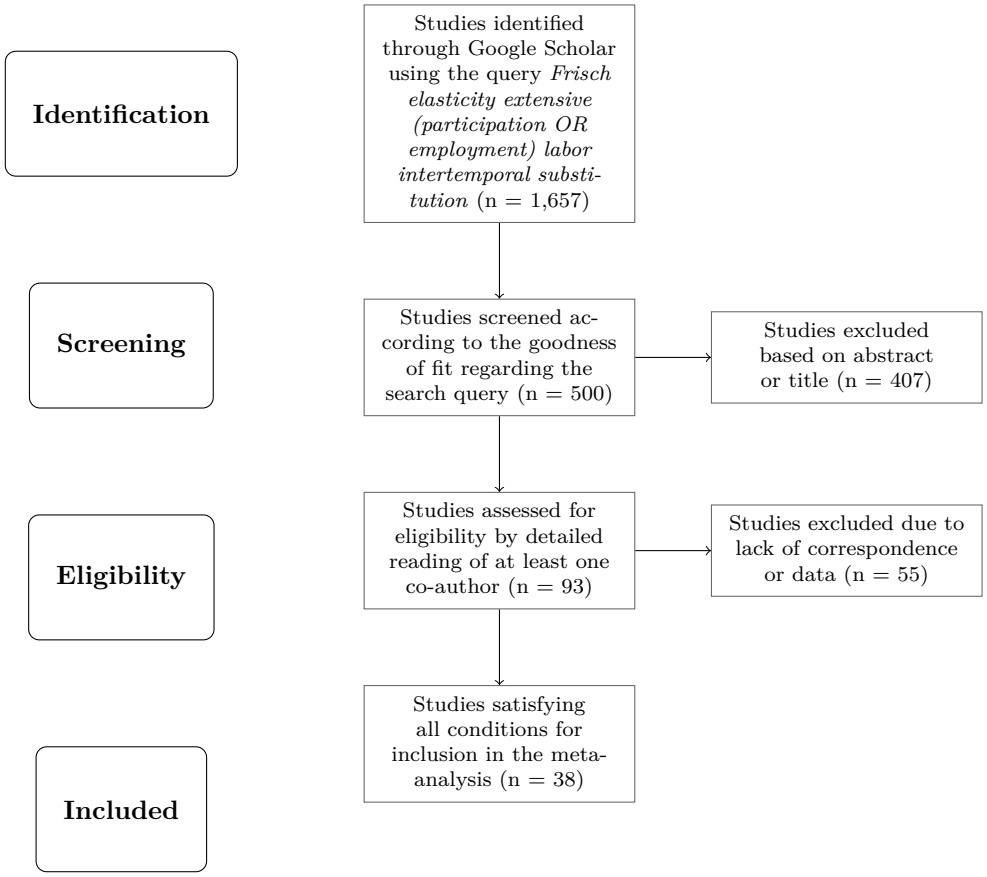
Figure 3.A11: Model size and convergence in the BMA (Random and HQ g-prior)



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 3.A7.

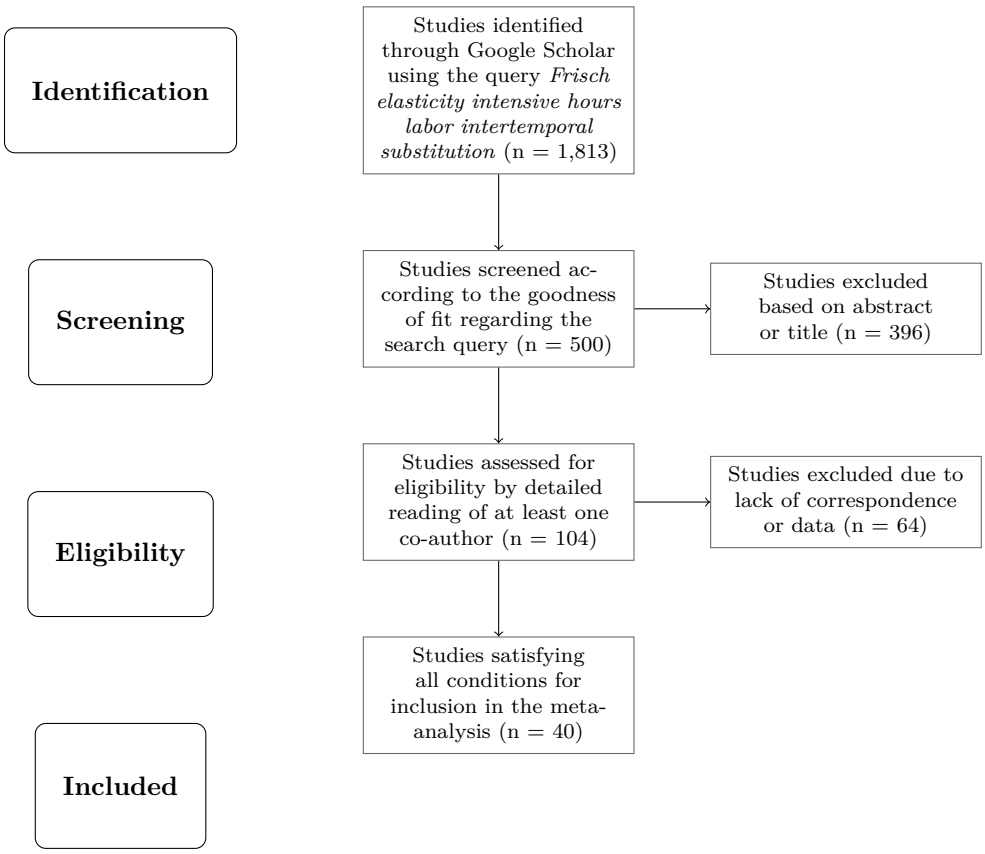
3.B Details on Literature Search and Data Collection

Figure 3.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 3.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 3.B1: Sources for estimates collected from individual papers

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

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

3.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 and Kim (2006), Keane (2011), and Attanasio et al. (2018). 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 total 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 total hours Frisch elasticity:

$$\eta = \frac{\partial h_t}{\partial w_t} \frac{w_t}{h_t} ||_{\lambda}, \tag{3.C1}$$

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

$$\ln h_t = \alpha_i + \rho + \theta x_t + \eta \ln w_t + \varepsilon_t, \tag{3.C2}$$

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

$$h_t = \begin{cases} \bar{h}, & \text{if } w_t \geq w_t^R \\ 0, & \text{if } w_t \leq w_t^R. \end{cases} \tag{3.C3}$$

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

$$\text{var}(\ln H_t) = \text{var}(\ln n_t) + \text{var}(\ln \bar{h}_t) + 2\,\text{cov}(\ln n_t, \ln \bar{h}_t), \tag{3.C4}$$

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 Equation 2.C4, the decomposition of total hours Frisch elasticity can be written as

$$\eta = \frac{\text{cov}(\Delta \ln H, \Delta \ln W)}{\text{var}(\Delta \ln W)} = \frac{\text{cov}(\Delta \ln \bar{h}, \Delta \ln W)}{\text{var}(\Delta \ln W)} + \frac{\text{cov}(\Delta \ln n, \Delta \ln W)}{\text{var}(\Delta \ln W)}, \tag{3.C5}$$

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 the main text.

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

Table 3.D1: 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 3.D2: Publication bias tests in a subsample of quasi-experimental estimates

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

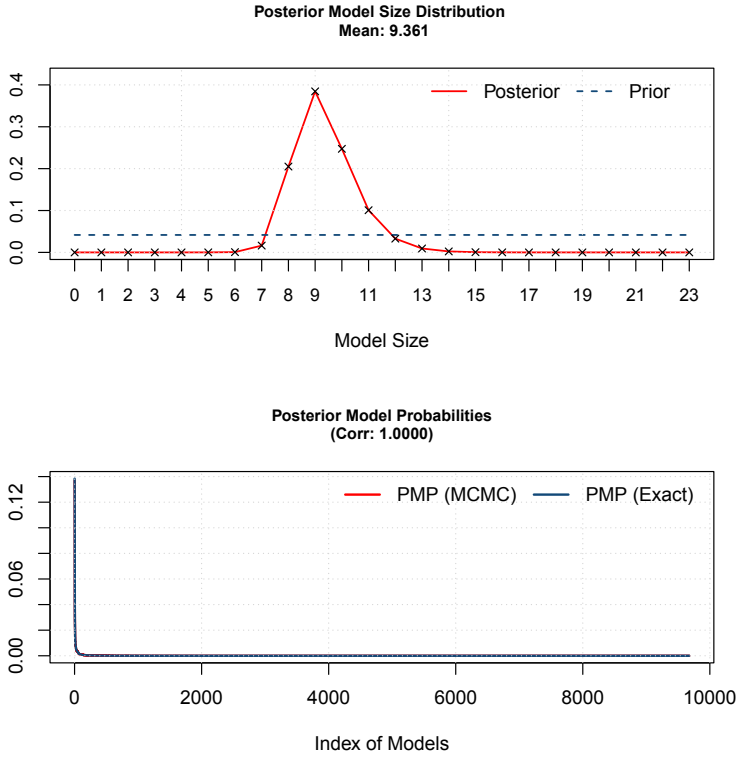
Notes: Panel A presents the results of regression $\hat{\eta}_{ij} = \eta_0 + \delta \cdot SE(\hat{\eta}_{ij}) + e_{ij}$, where $\hat{\eta}_{ij}$ and $SE(\hat{\eta}_{ij})$ are the i-th estimated Frisch extensive margin elasticity and its standard error reported in the j-th study. OLS = ordinary least 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 3.D3: Summary of the benchmark BMA estimation

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

Notes: 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 3.D1: Model size and convergence in the benchmark BMA model



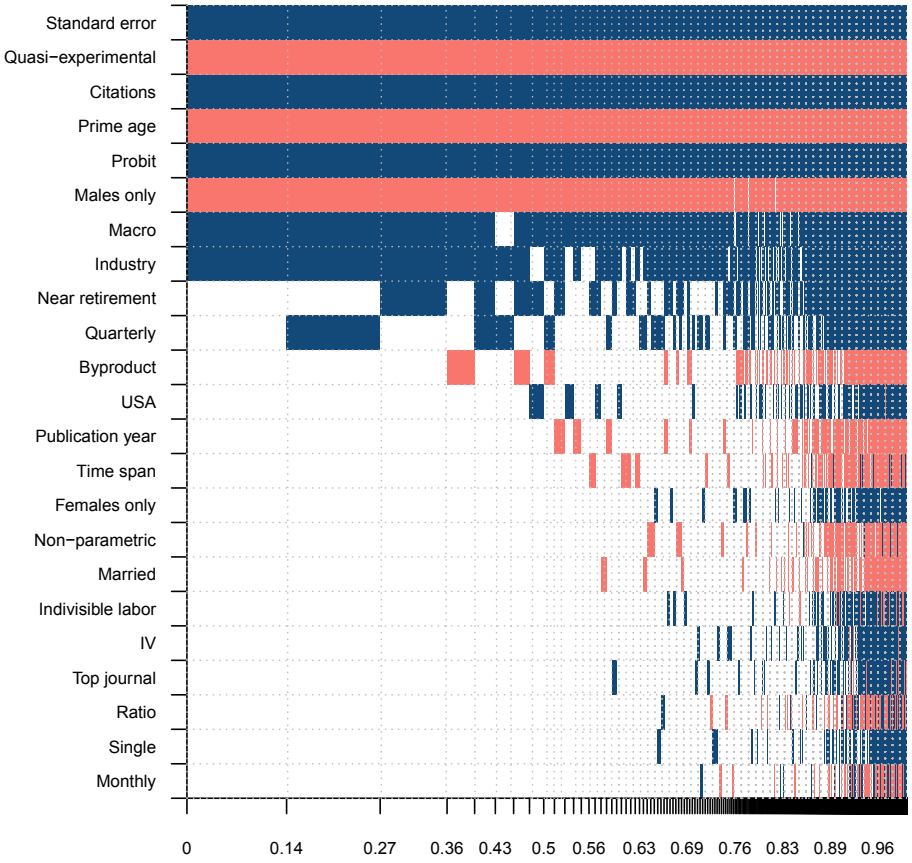
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA exercise reported in the main text.

Table 3.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
Demographic characteristics									
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
Data characteristics									
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
Specification characteristics									
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
Publication characteristics									
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. In the right-hand panel we use Mallow's weights Hansen (2007) and the orthogonalization of the covariate space suggested by Amini and Parmeter (2012) to conduct the frequentist model averaging exercise.

Figure 3.D2: Model inclusion in BMA (BRIC g-prior)



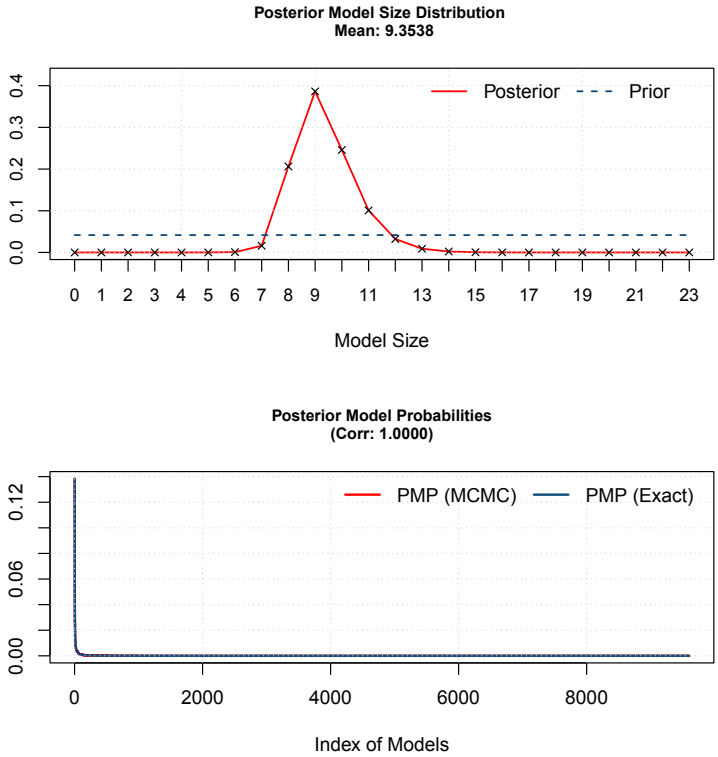
Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 3.D4.

Table 3.D5: Summary of the BMA (BRIC g-prior)

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

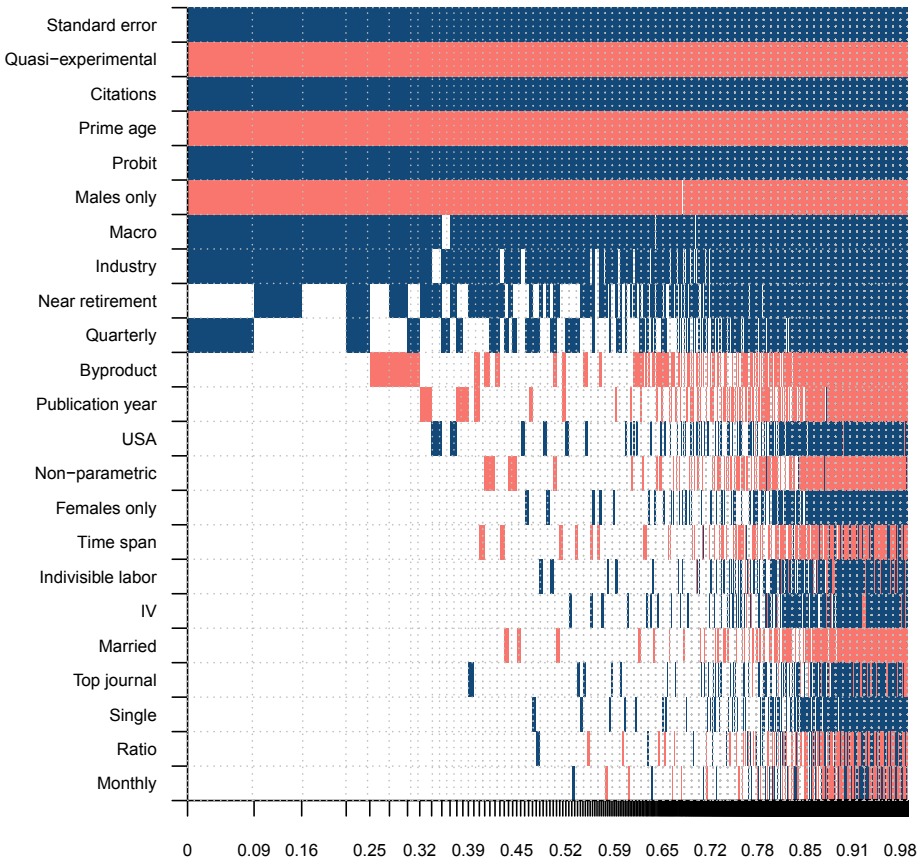
Notes: The results of this BMA specification are reported in Table 3.D4. The estimation is based on BRIC g-prior suggested by Fernandez et al. (2001) and the beta-binomial model prior according to Ley and Steel (2009).

Figure 3.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 3.D4.

Figure 3.D4: Model inclusion in BMA (Random and HQ g-prior)



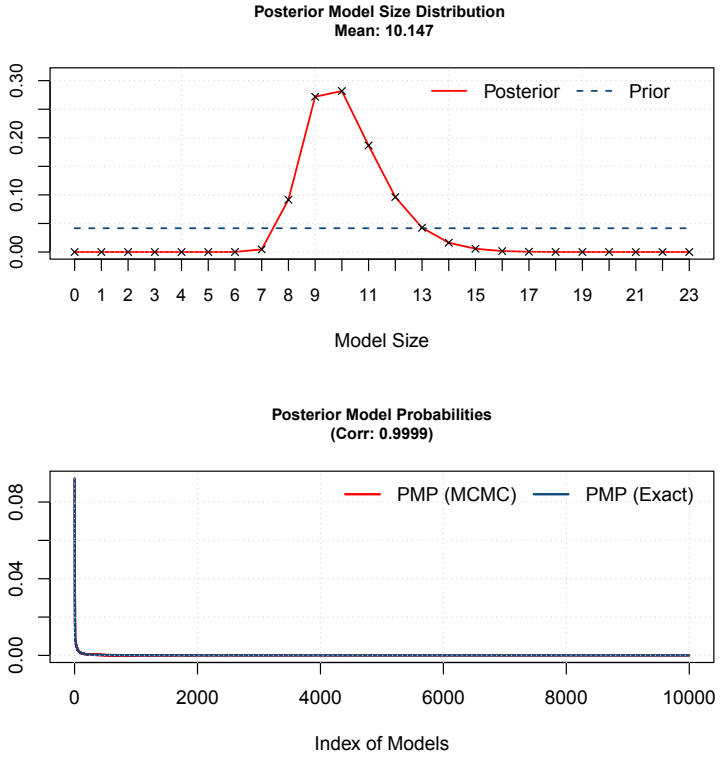
Notes: The response variable is the estimate of the Frisch extensive elasticity reported in a primary study. The columns denote individual models; variables are sorted by posterior inclusion probability in descending order. The horizontal axis denotes the cumulative posterior model probabilities. The 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 3.D4.

Table 3.D6: Summary of the BMA (Random and HQ g-prior)

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

Notes: The results of this BMA specification are reported in Table 3.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 3.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 3.D4.

Chapter 4

The Calvo Parameter Revisited: An Unbiased Insight

Abstract

This study provides a meta-analysis of the Calvo parameter estimated within the new Keynesian Phillips curve using a data set of 509 estimates from 40 studies published in a quarter century. Novel linear and nonlinear techniques suggest publication bias distorting the reported estimates towards typical values of the Calvo parameter used for calibration. Moreover, Bayesian model averaging results indicate that the reported estimates are systematically affected by various aspects of research design, particularly the choice of forcing variable in the NKPC, instrument selection, and authors' affiliation.

Keywords: Calvo parameter, New Keynesian Phillips Curve, meta-analysis, publication bias, Bayesian model averaging

JEL Codes: C11, C83, E31

This paper is published in Applied Economics Letters. I gratefully acknowledge financial support from the Czech Science Foundation grant "Spurious Precision in Meta-Analysis of Social Science Research" (#23 -05227M).

4.1 Introduction

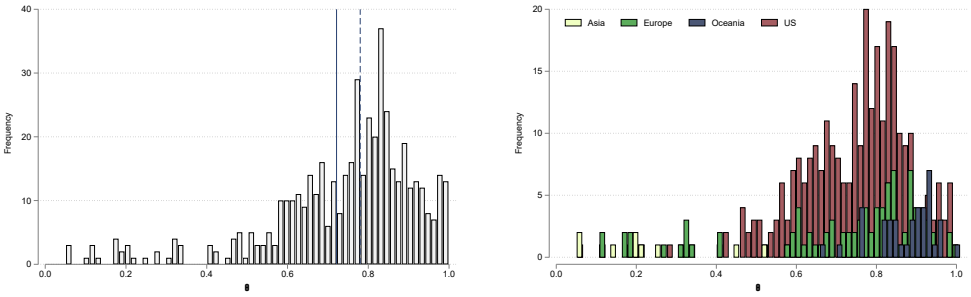
Standard Calvo-based New Keynesian Phillips Curve (NKPC) is one of the central components in dynamic macroeconomic modeling. According to Calvo (1983), there is a constant probability $(1-\theta)$ that in each period, a typical firm adjusts its price, and its price remains unchanged with probability θ, which is usually referred to as the Calvo parameter or price rigidity. Empirical examinations of the NKPC based on the Calvo pricing model result in a wide range of values. For example, Galí and Gertler (1999) and Galí et al. (2001) find price rigidity between 0.42 to 0.92 within the estimated NKPC. However, for calibration, researchers usually rely on the vast body of literature suggesting typical values such as 0.75 (average price duration of 4 quarters). For instance, Smets and Wouters (2003), as one of the influential studies in DSGE modeling, uses the typical value of 0.75 a priori for the Calvo parameter.

Two natural questions arise facing the Calvo parameter: First, is the parameter value consistent with the microeconomic data? Second, how does the estimated/calibrated parameter differ from the rest of the reported values in the literature? Extensive literature addresses the first question by comparing estimates based on the Calvo pricing model and microeconomic evidence. Alvarez and Burriel (2010) show that the standard Calvo model fails to capture the distribution of price durations found in microeconomic data. In contrast, Dufour et al. (2010) show that conditional on instrument selections, the price durations estimated by the Calvo-based NKPC are consistent with the US micro data. However, Nakamura and Steinsson (2013) argue that relying solely on the frequency of price changes might be misleading without considering other factors such as sales and cross-sectional heterogeneity. This paper mainly focuses on the second question by conducting a meta-study of a quarter-century literature. Meta-studies have become a widely accepted practice in economics

since they are crucial in explaining the variation of results between individual studies (see, e.g., Chetty et al. 2013; Gechert et al. 2022 and Havranek et al. 2022). Similarly, this paper studies how different sources of heterogeneity affect the Calvo parameter estimated within the structural NKPC. To do so, I use a dataset of 509 reported estimates from 40 studies over the last 24 years. The results obtained from various techniques imply the presence of publication bias in the literature. Furthermore, using the Bayesian averaging model (BMA), I show that choice of forcing variables, authors' affiliation, and a set of research characteristics systematically affect the estimates of the Calvo parameter. To my knowledge, this is the first meta-analysis investigating the sources of variation among estimated Calvo parameters.

4.2 Data

Figure 4.1: Patterns in the data



Notes: The solid line is the mean estimate, and the dashed line denotes the median estimate reported in primary studies. Outlier estimates (i.e., negative or larger than 1) are excluded from the sub-figures.

I use the Google Scholar search engine to find relevant estimates of the Calvo parameter in the literature. This database provides a powerful tool for full-text search. The Appendix 4.A provides details on the search process for collected estimates, which is consistent with the current protocol for meta-analysis (Havránek et al. 2020). The final dataset used in this paper covers 24 years of research from 1999 to 2022. It includes 509 estimates from 40 primary studies.

Table 4.1 lists the primary studies used in the meta-analysis. All collected parameters are estimated within the NKPC equation. The Calvo-based NKPC is typically given by:

$$\pi_t = \beta E[\pi_{t+1}] + \lambda mc_t, \tag{4.1}$$

where β is a subjective discount factor, mc_t is real marginal costs, and $\lambda = (1-\theta)(1-\beta\theta)/\theta$. Hence, collected values of θ are obtained by structural estimates of the NKPC. Moreover, estimates are collected with their corresponding standard errors. Therefore, estimates reported without standard errors are excluded from the dataset. In addition to reported estimates and standard errors, the dataset includes 26 extra explanatory variables reflecting the framework in which estimates are reported: data characteristics, model specifications, estimation techniques, and publication characteristics. The final dataset consists of more than 14,000 manually collected data points. The search termination date is December 31, 2022. As of the search termination date, all the studies received 9679 citations combined, indicating the importance of primary studies. Table 4.C1 provides more details on explanatory variables.

The left-hand side of Figure 4.1 illustrates the distribution of reported estimates in the literature. Overall, estimates are concentrated mainly around 0.70 and 0.85. However, there are several outliers on both sides of the distribution. Therefore, the data are winsorized at the 5% level. The mean point estimate is 0.72 (solid line), marginally smaller than the median, 0.78 (dashed line). The mean estimate is slightly lower than the typical value (0.75) used in the calibration of the dynamic model. The right-hand side of Figure 4.1 pictures the differences in the distribution of estimates from different regions. There is substantial variation among estimates if we consider different regions, which are not necessarily consistent with the microeconomic data. In the case of the US, the mean estimate and implied average price duration are 0.74 and

Table 4.1: Studies used in the meta-analysis

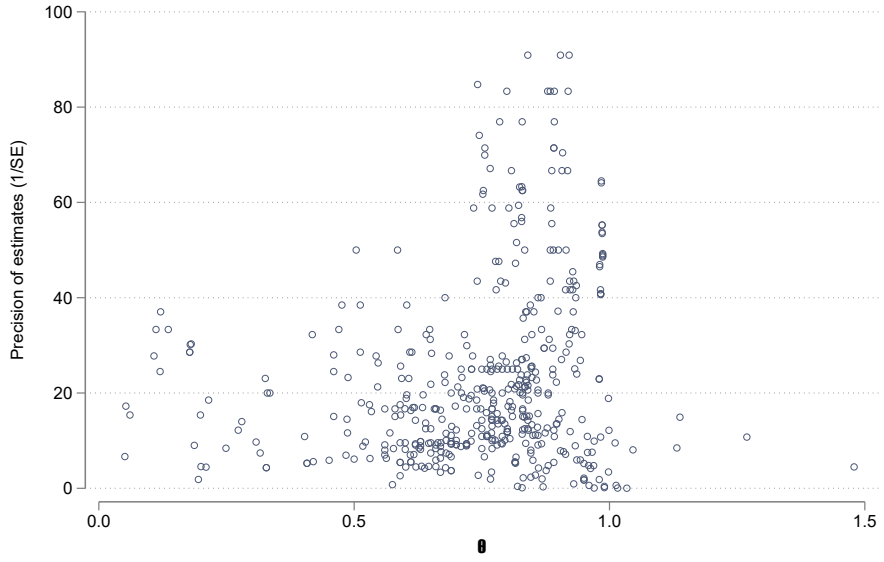
Abbas (2022)	Galí et al. (2001)
Abbas and Sgro (2011)	Guerrieri et al. (2010)
Adam and Padula (2011)	Hung and Kwan (2022)
Ahrens and Sacht (2014)	Kurachi et al. (2016)
Alvarez and Burriel (2010)	Kurmann (2007)
Arslan (2010)	Kuttner and Robinson (2010)
Ascari and Sbordone (2014)	Lawless and Whelan (2011)
Berardi and Galimberti (2017)	Lie and Yadav (2017)
Celasun (2006)	Madeira (2014)
Chin (2019)	Martins and Gabriel (2009)
Christensen and Dib (2008)	Matheron and Maury (2004)
Fiore and Tristani (2013)	McAdam and Willman (2004)
Walque et al. (2006)	Muscatelli et al. (2004)
Dib (2011)	Nunes (2010)
Dufour et al. (2010)	Ravenna and Walsh (2006)
Eichenbaum and Fisher (2007)	Scheufele (2010)
Furuoka et al. (2020)	Sheedy (2010)
Furuoka et al. (2021)	Smets and Wouters (2002)
Gabriel and Martins (2010)	Vázquez et al. (2012)
Galí and Gertler (1999)	Yazgan and Yilmazkuday (2005)

3.8 quarters, respectively. These numbers are in agreement with the part of the empirical literature on the US data (Nakamura and Steinsson 2013; Cravino et al. 2020. Additionally, the mean point estimate based on European data is 0.70, which implies an average price duration of around ten months. This value is inconsistent with some of the microeconomic evidence from the euro area (Alvarez et al. 2006). However, this mean estimate is in line with more recent microeconomic evidence from the euro area. For example, in a recent study, Gautier et al. (2022) find an average price duration in 11 countries in the euro area between 3.39 and 5.15 quarters, depending on the inclusion of sales in the data, which is partially consistent with the mean estimate of European data.

4.3 Publication Bias

Publication bias significantly affects reported estimates in different fields of science, including economics. Researchers systematically tend to report estimates that are statistically significant and avoid estimates that are either insignificant or with a wrong sign. Hence, one can interpret the relationship between reported estimates and their standard errors as publication bias. Figure 4.2 shows the relationship between the reported Calvo parameters and their corresponding precision (the inverse of standard error). Without publication bias, the scatter plot (funnel plot) should form a symmetric inverted funnel since the most precise estimates would be around the average effect, and the estimates with lower precision would be more dispersed. Therefore, since there is a noticeable asymmetry, this visual tool suggests the presence of publication bias in the literature.

Figure 4.2: Funnel plot suggests publication bias



Notes: In the absence of publication bias, the plot should resemble a symmetric inverted funnel. Outliers are excluded from the figure but included in the analysis.

Relying solely on a visual tool in which we assume a linear relationship be-

tween estimates and their precision is insufficient to conclude publication bias. Regressing estimates against their standard errors, one can extend assessing the asymmetry of the funnel plot to a regression-based test:

$$\hat{\theta}_{ij} = \theta_0 + \beta \cdot SE(\hat{\theta}_{ij}) + \epsilon_{ij}, \tag{4.2}$$

where $\hat{\theta}_{ij}$ is the i^{th} reported estimate in the j^{th} study and $SE(\hat{\theta}_{ij})$ is the corresponding standard error. In this regression setting, β denotes the size of publication bias, and the intercept can be interpreted as the mean value of the estimate corrected for publication bias. Panel A in Table 4.2 reports the regression results based on different specifications. Since the original regression is subject to heteroskedasticity, both sides of Equation 4.3 are divided by standard errors to give more weights to more precise estimates, which yields a weighted least squares estimator. Besides, standard errors are clustered at the study level since estimates within a study are not independent. The weighted estimator and additional specifications (except the study fixed effect specification) imply publication bias in estimating the Calvo parameter. Furthermore, the results indicate that if we exclude systematic publication bias, the mean corrected for bias will vary between 0.75 and 0.90, depending on the specification. This variation means that the average price rigidity can be 5% to 25% larger, which consequently implies an average price duration of up to twice a longer period (8 quarters) than what the mean estimate in the literature suggests.

The formal linear tests assume a robust linear relationship between the reported estimates and the standard errors. However, several studies argue that this relationship is not necessarily linear (Andrews and Kasy 2019). Relaxing this assumption, I use four nonlinear techniques to investigate publication bias. These methods usually assume that the linear correlation between the effect

Table 4.2: Linear and nonlinear tests

Panel A: Linear tests	WLS	FE	BE	Study
Standard error	−1.749***	0.380	−3.289***	−2.205***
(*publication bias*)	(0.503)	(0.916)	(0.913)	(0.529)
	[-2.986, -0.735]			[-3.310, -1.130]
Constant	0.840***	0.754***	0.879***	0.842***
(*mean beyond bias*)	(0.022)	(0.037)	(0.025)	(0.0187)
	[0.790, 0.887]			[0.801, 0.881]
Implied duration (quarters)	6.250	4.065	8.264	6.329
Observations	509	509	509	509
Studies	40	40	40	40
Panel B: Nonlinear tests	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.800***	0.785***	0.800***	0.819***
	(0.022)	(0.009)	(0.007)	(0.030)
Implied duration (quarters)	5.000	4.651	5.000	5.525
Observations	509	509	509	509
Studies	40	40	40	40

Notes: Panel A presents the results of Equation 4.3. WLS = weighted least squares. FE = study fixed effects. Study = estimates are weighted by the inverse of the number of estimates reported per study. Standard errors are clustered at the study level; 95% confidence intervals from wild bootstrap clustering are reported in square brackets, if applicable. Panel B presents the mean effect corrected for publication bias using nonlinear techniques. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

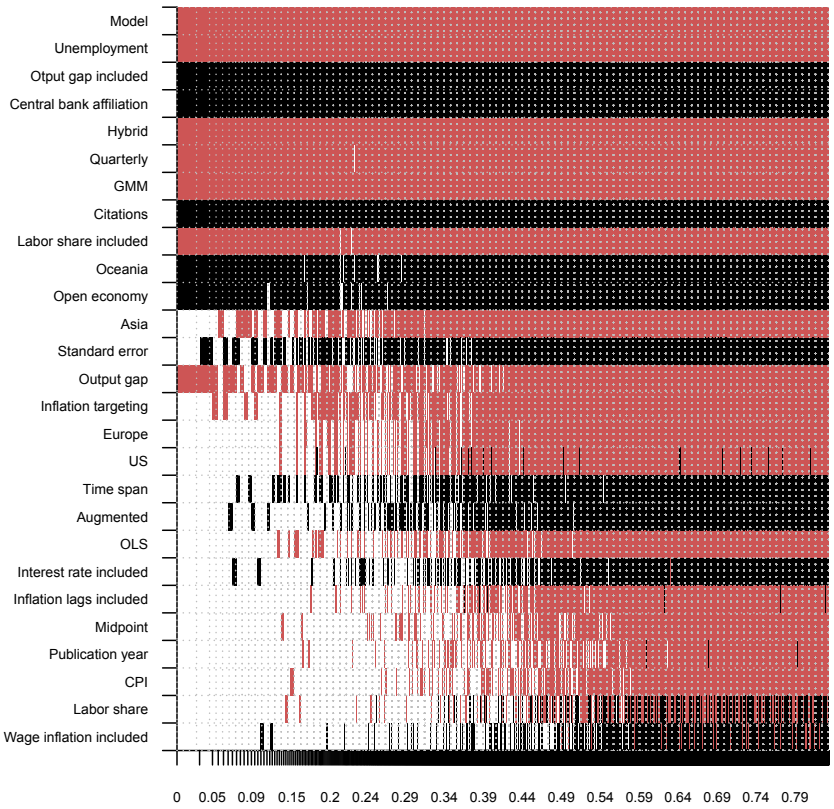
and its standard error is distorted by crossing different precision thresholds. Panel B in Table 4.2 reports the results of the different nonlinear techniques. The results are consistent with linear regressions, as they yield a mean beyond bias (between 0.76 and 0.82) larger than the mean reported estimate in the literature. Similar to these results, Meenagh et al. (2022) show that in the case of Bayesian methods, estimates of price rigidity are biased toward the adopted priors, which are usually close to the mean (common) estimate in the literature. The appendices provides details on nonlinear methods and additional results from different subsamples for both linear and nonlinear techniques.

4.4 Heterogeneity

The first set of results indicates the effect of publication bias on estimates. However, publication bias may be the product of heterogeneity among estimates. To address heterogeneity, 26 additional explanatory variables are used

that reflect various aspects of studies in which estimates are reported. The Appendix 3.C provides more details about explanatory variables.

Figure 4.3: Model inclusion in Bayesian model averaging



Notes: The response variable is the Calvo parameter estimated within the NKPC. 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 estimate is based on the unit information prior (UIP) recommended by Eicher et al. (2011) and the dilution prior suggested by George (2010), which takes into account collinearity. Black (darker in grayscale) = the variable has a positive estimated sign. Red (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 4.3.

The first option for investigating heterogeneity is a simple OLS test to regress the reported estimates on the set of explanatory variables. This simple regression, however, does not address the issue of model uncertainty. To this end, I use Bayesian model averaging (BMA) to capture model uncertainty. Using various subsets of explanatory variables, BMA runs multiple regressions and ranks models' relative performance by their posterior model probabilities

(PMP). Moreover, posterior inclusion probability (PIP) for each variable indicates the sum of posterior model probabilities (PMP) of the models in which the variable is included. Since, in our case, the number of models visited by BMA is significantly large (2^{27}), I apply the birth-and-death Markov chain Monte Carlo (BDMCMC) algorithm proposed by Stephens (2000), which includes models with the highest PMP. I use the `bms` package developed by Zeugner and Feldkircher (2015). Considering the collinearity of the variables included in each model, I employ the dilution prior, suggested by George (2010), in the benchmark specification. The appendices provide more details on BMA methods. Moreover, based on the benchmark BMA results, I run a frequentist OLS check, including only variables with PIPs higher than 0.5 obtained from the benchmark BMA specification. Lastly, using Mallow's weights (Hansen 2007) and the orthogonalization of covariate space (Amini and Parmeter 2012), I apply frequentist model averaging (FMA), which assumes that explanatory variables are fixed and does not rely on probabilistic information based on prior knowledge. In a comprehensive and insightful study, Steel (2020) discusses the BMA and FMA methods in greater detail.

Figure 4.3 illustrates the results of the benchmark BMA specification. In addition, the left-hand panel in Table 4.3 reports the corresponding numerical result. Based on these results, fourteen variables with a PIP larger than 0.5 systematically affect the size of the estimates. Among data characteristics, not surprisingly, the region in which data is taken significantly affects the size of estimated parameters. In addition, studies conducted in countries with an inflation-targeting monetary policy tend to report smaller parameter values. As an intuitive result, model specifications significantly impact reported estimates. Although the choice of inflation measure seems not to have a systematic effect on the estimated parameters, the magnitude of estimates is sensitive to the choice of forcing variable.

Table 4.3: Explaining heterogeneity

	BMA			OLS			FMA		
Variable	Post. Mean	Post. SD	PIP	Coeff.	S.E.	P-val.	Coeff.	S.E.	P-val.
Constant	0.989	N.A.	1.000	0.941	0.078	0.000	1.127	0.098	0.000
Standard error	0.143	0.141	0.595	0.217	0.269	0.426	0.294	0.102	0.004
Data characteristics									
Time span	0.017	0.027	0.365				0.035	0.024	0.156
Midpoint	-0.002	0.006	0.171				-0.013	0.011	0.218
Quarterly	-0.205	0.063	0.988	-0.197	0.078	0.015	-0.233	0.059	0.000
Inflation targeting	-0.045	0.053	0.511	-0.060	0.047	0.215	-0.101	0.038	0.007
US	-0.075	0.100	0.444				-0.183	0.053	0.001
Europe	-0.079	0.102	0.467				-0.193	0.054	0.000
Oceania	0.143	0.085	0.822	0.233	0.069	0.002	0.095	0.055	0.088
Asia	-0.114	0.109	0.652	-0.057	0.086	0.509	-0.184	0.063	0.003
Specifications									
Hybrid	-0.065	0.020	0.984	-0.074	0.031	0.022	-0.056	0.018	0.002
Open economy	0.068	0.043	0.806	0.098	0.042	0.027	0.076	0.033	0.021
Model	-0.177	0.033	1.000	-0.159	0.079	0.052	-0.178	0.035	0.000
Augmented	0.021	0.033	0.382				0.059	0.026	0.023
CPI	-0.003	0.011	0.166				-0.015	0.020	0.450
Labor share	0.000	0.020	0.163				-0.019	0.043	0.661
Unemployment	-0.361	0.065	1.000	-0.346	0.119	0.006	-0.386	0.068	0.000
Output gap	-0.045	0.053	0.520	-0.058	0.081	0.476	-0.053	0.057	0.355
Estimation techniques									
OLS	-0.023	0.045	0.293				-0.091	0.049	0.063
GMM	-0.092	0.032	0.967	-0.097	0.041	0.023	-0.089	0.030	0.003
Inflation lags included	-0.021	0.047	0.254				-0.100	0.049	0.041
Labor share included	-0.095	0.033	0.957	-0.108	0.032	0.002	-0.072	0.028	0.010
Output gap included	0.120	0.027	0.999	0.136	0.037	0.001	0.113	0.026	0.000
Interest rate included	0.010	0.024	0.243				0.017	0.031	0.586
Wage inflation included	0.002	0.011	0.142				0.009	0.024	0.713
Publication characteristics									
Publication year	-0.004	0.012	0.177				-0.003	0.019	0.891
Central bank affiliation	0.096	0.026	0.997	0.091	0.041	0.032	0.090	0.029	0.002
Citations	0.028	0.010	0.946	0.028	0.014	0.046	0.034	0.011	0.002
Observations	509			509			509		
Studies	40			40			40		

Notes: The response variable is the Calvo parameter estimated within the NKPC. SD = standard deviation, PIP = posterior inclusion probability, SE = standard error. The left panel applies BMA based on the UIP g prior and the dilution prior (Eicher et al. 2011; George 2010). The middle panel reports a frequentist check using OLS, which includes variables with PIPs greater than 0.50 in the benchmark BMA. Standard errors in the frequentist check are clustered at the study level. To conduct the frequentist model averaging, reported on the right panel, we use Mallow's weights by Hansen (2007) and the orthogonalization of the covariate space suggested by Amini and Parmeter (2012). Table 4.C1 presents a detailed description of all variables.

Unemployment and output gaps are associated with lower values of the Calvo parameter. Furthermore, BMA results indicate that using the GMM estimator to account for endogeneity could systematically result in smaller estimates. Similar to the forcing variable, the output gap among instruments is systematically associated with smaller estimates. The results also suggest that higher citations are associated with larger estimates. Likewise, studies with at least one author affiliated with a central bank tend to report larger estimates. Finally, the results of frequentist OLS and FMA checks are generally consistent with the benchmark BMA findings. In addition to crucial variables highlighted in the Bayesian setting, FMA results suggest that the estimates are sensitive to all regional data as well as to the OLS method. More details and robustness checks are provided in the appendices.

4.5 Conclusion

This paper provides a meta-analysis of the literature on Calvo-based NKPC estimates. The results based on a dataset of 509 reported estimates from 40 studies suggest that publication bias is present in the literature, distorting the reported estimates towards more orthodox values of the Calvo parameter. The linear and nonlinear techniques suggest that the implied average price durations, after correcting for publication bias, exhibit some discrepancies with microeconomic data evidence.

Moreover, the benchmark Bayesian model averaging results show that model specifications, in particular the choice of forcing variables in the NKPC, play a significant role in determining the Calvo parameter value. Surprisingly, no evidence indicates that the inflation measure is systematically correlated with the magnitude of estimates. Similarly, the estimated parameters are sensitive to instrument selection. Finally, in addition to using quarterly data, the cen-

tral bank's inflation targeting strategy tends to shrink the value of estimated parameters. On the other hand, the results indicate that central bank affiliation is positively associated with larger estimates of the Calvo parameter. Robustness checks are also in line with the findings of the benchmark BMA setting. These results provide a complementary set of helpful information to calibrate and estimate the Calvo parameter. Researchers may use an unbiased Calvo parameter to calibrate within the empirical Calvo-based NKPC, based on the context of their research (e.g., the choice of proxy for marginal costs and the region where data are obtained). Similarly, the results are helpful for comparative analyzes of the estimated NKPC. Further studies can extend the framework of this paper by investigating other aspects of research design to estimate the Calvo parameter absent from this paper.

References

Abbas, S. K. (2022). Asymmetry in the regimes of inflation and business cycles: the new Keynesian Phillips curve. *Applied Economics*, pages 1–14.

Abbas, S. K. and Sgro, P. M. (2011). New Keynesian Phillips curve and inflation dynamics in australia. *Economic Modelling*, 28(4):2022–2033.

Adam, K. and Padula, M. (2011). Inflation dynamics and subjective expectations in the United States. *Economic Inquiry*, 49(1):13–25.

Ahrens, S. and Sacht, S. (2014). Estimating a high-frequency new-Keynesian Phillips curve. *Empirical Economics*, 46(2):607–628.

Alvarez, L. J. and Burriel, P. (2010). Is a Calvo price setting model consistent with individual price data? *The BE Journal of Macroeconomics*, 10(1).

Alvarez, L. J., Dhyne, E., Hoeberichts, M., Kwapil, C., Le Bihan, H., Lünne-

mann, P., Martins, F., Sabbatini, R., Stahl, H., Vermeulen, P., et al. (2006). Sticky prices in the Euro area: a summary of new micro-evidence. *Journal of the European Economic Association*, 4(2-3):575–584.

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

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

Arslan, M. M. (2010). Relative importance of sticky prices and sticky information in price setting. *Economic Modelling*, 27(5):1124–1135.

Ascari, G. and Sbordone, A. M. (2014). The macroeconomics of trend inflation. *Journal of Economic Literature*, 52(3):679–739.

Berardi, M. and Galimberti, J. K. (2017). On the initialization of adaptive learning in macroeconomic models. *Journal of Economic Dynamics and Control*, 78:26–53.

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

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12(3):383–398.

Celasun, O. (2006). Sticky inflation and the real effects of exchange rate-based stabilization. *Journal of International Economics*, 70(1):115–139.

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

Chin, K.-H. (2019). New Keynesian Phillips curve with time-varying parameters. *Empirical Economics*, 57(6):1869–1889.

Christensen, I. and Dib, A. (2008). The financial accelerator in an estimated new Keynesian model. *Review of economic dynamics*, 11(1):155–178.

Cravino, J., Lan, T., and Levchenko, A. A. (2020). Price stickiness along the income distribution and the effects of monetary policy. *Journal of Monetary Economics*, 110:19–32.

Dib, A. (2011). Monetary policy in estimated models of small open and closed economies. *Open Economies Review*, 22(5):769–796.

Dufour, J.-M., Khalaf, L., and Kichian, M. (2010). On the precision of Calvo parameter estimates in structural NKPC models. *Journal of Economic Dynamics and Control*, 34(9):1582–1595.

Eichenbaum, M. and Fisher, J. D. (2007). Estimating the frequency of price re-optimization in Calvo-style models. *Journal of monetary Economics*, 54(7):2032–2047.

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

Fiore, F. D. and Tristani, O. (2013). Optimal monetary policy in a model of the credit channel. *The Economic Journal*, 123(571):906–931.

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

Furuoka, F., Ling, P. K., Chomar, M. T., and Nikitina, L. (2020). Trade openness and the Phillips curve: Evidence from asean countries. *The Singapore Economic Review*, pages 1–25.

Furuoka, F., Pui, K. L., Chomar, M. T., and Nikitina, L. (2021). Is the Phillips curve disappearing? evidence from a new test procedure. *Applied Economics Letters*, 28(6):493–500.

Gabriel, V. J. and Martins, L. F. (2010). The cost channel reconsidered: A comment using an identification-robust approach. *Journal of Money, Credit and Banking*, 42(8):1703–1712.

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

Galí, J., Gertler, M., and Lopez-Salido, J. D. (2001). European inflation dynamics. *European Economic Review*, 45(7):1237–1270.

Gautier, E., Conflitti, C., Faber, R. P., Fabo, B., Fadejeva, L., Jouvanceau, V., Menz, J.-O., Messner, T., Petroulas, P., Roldan-Blanco, P., et al. (2022). New facts on consumer price rigidity in the euro area.

Gechert, S., Havranek, T., Irsova, Z., and Kolcunova, D. (2022). Measuring capital-labor substitution: The importance of method choices and publication bias. *Review of Economic Dynamics*, 45:55–82.

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

Guerrieri, L., Gust, C., and López-Salido, J. D. (2010). International competition and inflation: a new Keynesian perspective. *American Economic Journal: Macroeconomics*, 2(4):247–80.

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

Havranek, T., Irsova, Z., Laslopova, L., and Zeynalova, O. (2022). Publication and attenuation biases in measuring skill substitution. *The Review of Economics and Statistics*, 1(1):1–37.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and Van Aert, R. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, 34(3):469–475.

Hung, T. H. and Kwan, Y. K. (2022). Hong Kong's new Keynesian Phillips curve: Sticky information or sticky price? *Pacific Economic Review*, 27(1):42–55.

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

Kurachi, Y., Hiraki, K., Nishioka, S., et al. (2016). Does a higher frequency of micro-level price changes matter for macro price stickiness?: Assessing the impact of temporary price changes. Technical report, Bank of Japan.

Kurmann, A. (2007). Var-based estimation of euler equations with an application to new Keynesian pricing. *Journal of Economic Dynamics and Control*, 31(3):767–796.

Kuttner, K. and Robinson, T. (2010). Understanding the flattening Phillips curve. *The North American Journal of Economics and Finance*, 21(2):110–125.

Lawless, M. and Whelan, K. T. (2011). Understanding the dynamics of labor shares and inflation. *Journal of Macroeconomics*, 33(2):121–136.

Lie, D. and Yadav, A. S. (2017). Time-varying trend inflation and the new Keynesian Phillips curve in australia. *Economic Record*, 93(300):42–66.

Madeira, J. (2014). Overtime labor, employment frictions, and the new Keynesian Phillips curve. *Review of Economics and Statistics*, 96(4):767–778.

Martins, L. F. and Gabriel, V. J. (2009). New Keynesian Phillips curves and potential identification failures: A generalized empirical likelihood analysis. *Journal of Macroeconomics*, 31(4):561–571.

Matheron, J. and Maury, T.-P. (2004). Supply-side refinements and the new Keynesian Phillips curve. *Economics Letters*, 82(3):391–396.

Mavroeidis, S., Plagborg-Møller, M., and Stock, J. H. (2014). Empirical evidence on inflation expectations in the new Keynesian Phillips curve. *Journal of Economic Literature*, 52(1):124–88.

McAdam, P. and Willman, A. (2004). Supply, factor shares and inflation persistence: Re-examining euro-area new-Keynesian Phillips curves. *Oxford Bulletin of Economics and Statistics*, 66:637–670.

Meenagh, D., Minford, P., and Wickens, M. R. (2022). The macroeconomic controversy over price rigidity—how to resolve it and how bayesian estimation has led us astray. *Open Economies Review*, pages 1–14.

Muscatelli, V. A., Tirelli, P., and Trecroci, C. (2004). Fiscal and monetary policy interactions: Empirical evidence and optimal policy using a structural new-Keynesian model. *Journal of Macroeconomics*, 26(2):257–280.

Nakamura, E. and Steinsson, J. (2013). Price rigidity: Microeconomic evidence and macroeconomic implications. *Annual Review of Economics*, 5(1):133–163.

Nunes, R. (2010). Inflation dynamics: the role of expectations. *Journal of Money, Credit and Banking*, 42(6):1161–1172.

Ravenna, F. and Walsh, C. E. (2006). Optimal monetary policy with the cost channel. *Journal of Monetary Economics*, 53(2):199–216.

Scheufele, R. (2010). Evaluating the german (new Keynesian) Phillips curve. *The North American Journal of Economics and Finance*, 21(2):145–164.

Sheedy, K. D. (2010). Intrinsic inflation persistence. *Journal of Monetary Economics*, 57(8):1049–1061.

Smets, F. and Wouters, R. (2002). Openness, imperfect exchange rate pass-through and monetary policy. *Journal of Monetary Economics*, 49(5):947–981.

Smets, F. and Wouters, R. (2003). An estimated dynamic stochastic general equilibrium model of the Euro area. *Journal of the European economic association*, 1(5):1123–1175.

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

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74.

Vázquez, J., María-Dolores, R., and Londoño, J. M. (2012). The effect of data revisions on the basic new Keynesian model. *International Review of Economics & Finance*, 24:235–249.

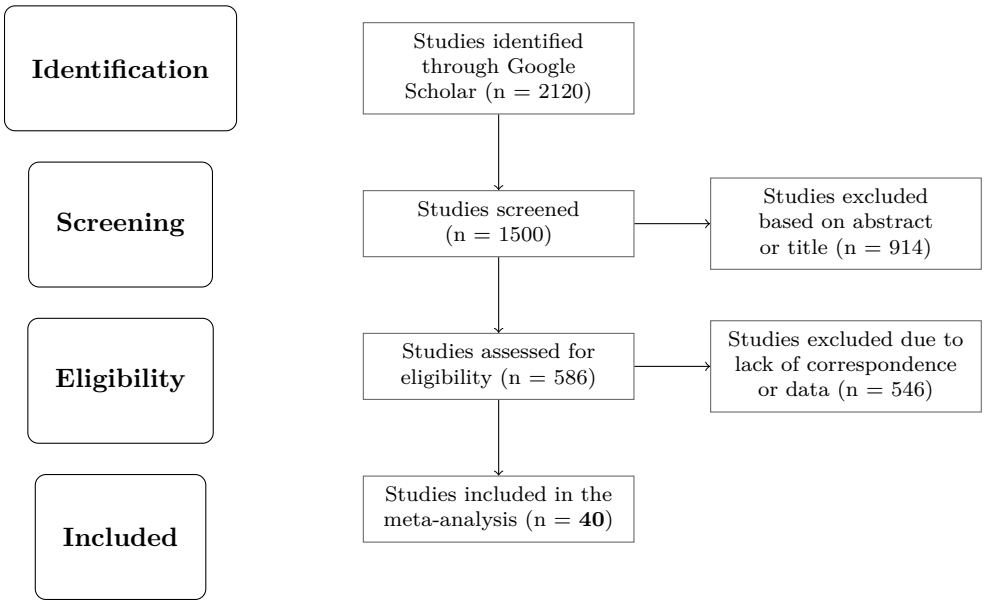
Walque, G. d., Smets, F., and Wouters, R. (2006). Price shocks in general equilibrium: Alternative specifications. *CESifo Economic Studies*, 52(1):153–176.

Yazgan, M. E. and Yilmazkuday, H. (2005). Inflation dynamics of turkey: a structural estimation. *Studies in Nonlinear Dynamics & Econometrics*, 9(1).

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

4.A Literature Search

Figure 4.A1: PRISMA flow diagram

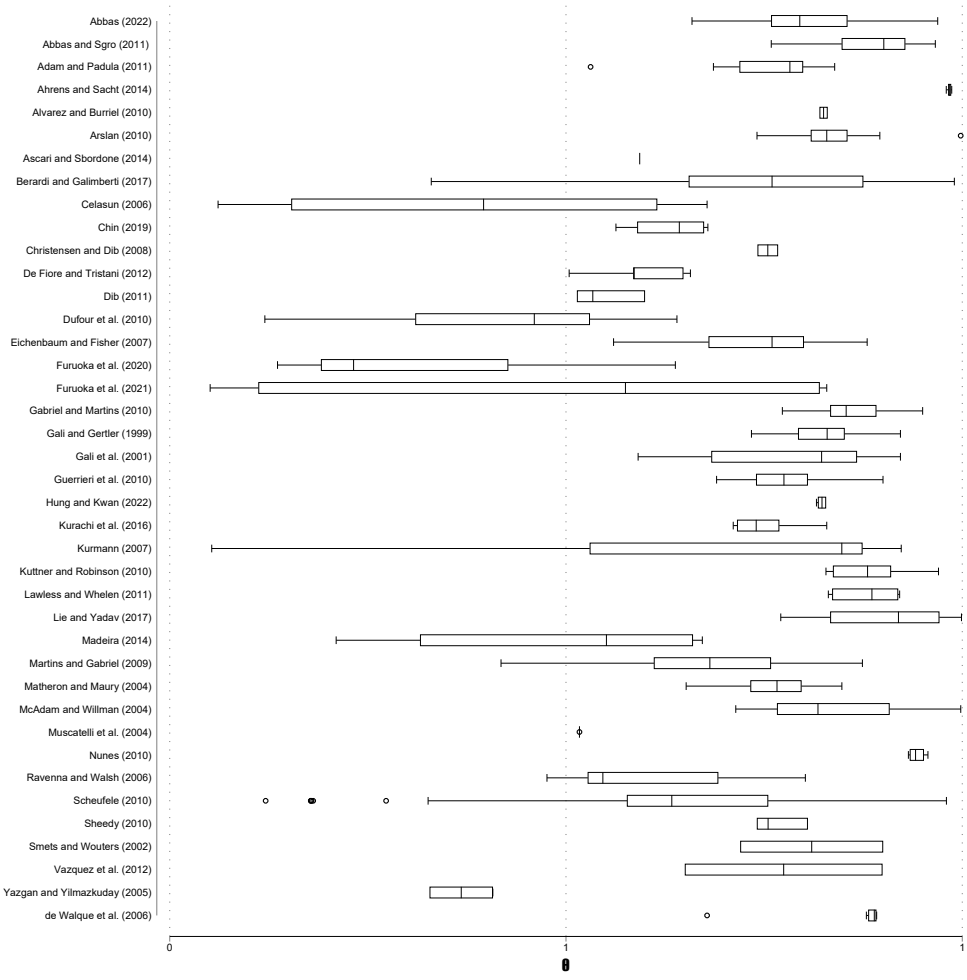


Notes: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) is an evidence-based set of items for reporting in systematic reviews and meta-analyses. More details on PRISMA and reporting standards of meta-analysis in general are provided by Havránek et al. (2020).

Table 4.A1: Studies used in the meta-analysis

Abbas (2022)	Galí et al. (2001)
Abbas and Sgro (2011)	Guerrieri et al. (2010)
Adam and Padula (2011)	Hung and Kwan (2022)
Ahrens and Sacht (2014)	Kurachi et al. (2016)
Alvarez and Burriel (2010)	Kurmann (2007)
Arslan (2010)	Kuttner and Robinson (2010)
Ascari and Sbordone (2014)	Lawless and Whelan (2011)
Berardi and Galimberti (2017)	Lie and Yadav (2017)
Celasun (2006)	Madeira (2014)
Chin (2019)	Martins and Gabriel (2009)
Christensen and Dib (2008)	Matheron and Maury (2004)
Fiore and Tristani (2013)	McAdam and Willman (2004)
Walque et al. (2006)	Muscatelli et al. (2004)
Dib (2011)	Nunes (2010)
Dufour et al. (2010)	Ravenna and Walsh (2006)
Eichenbaum and Fisher (2007)	Scheufele (2010)
Furuoka et al. (2020)	Sheedy (2010)
Furuoka et al. (2021)	Smets and Wouters (2002)
Gabriel and Martins (2010)	Vázquez et al. (2012)
Galí and Gertler (1999)	Yazgan and Yilmazkuday (2005)

Figure 4.A2: Variation of the estimates within and between studies



4.B Additional results for publication bias

4.B.1 Linear tests

Table 4.B1: Linear funnel asymmetry tests: GDP deflator and CPI

Panel A: GDP deflator	WLS	FE	BE	Study
Standard error	-1.572^{***}	-0.023	-3.356^{***}	-2.161^{***}
(*publication bias*)	(0.566)	(1.018)	(1.126)	(0.630)
	[-2.893, -0.216]			[-3.512, -0.839]
Constant	0.829^{***}	0.771^{***}	0.878^{***}	0.836^{***}
(*mean beyond bias*)	(0.020)	(0.038)	(0.033)	(0.027)
	[0.775, 0.868]			[0.768, 0.897]
Observations	353	353	353	353
Studies	28	28	28	28
Panel B: CPI	WLS	FE	BE	Study
Standard error	-2.395^{**}	2.406	-2.868	-2.214^{**}
(*publication bias*)	(1.054)	(1.977)	(1.630)	(1.059)
	[-6.172, 0.141]			[-4.738, 0.362]
Constant	0.886^{***}	0.647^{***}	0.876^{***}	0.852^{***}
(*mean beyond bias*)	(0.077)	(0.099)	(0.043)	(0.023)
	[0.625, 1.08]			[0.791, 0.974]
Observations	156	156	156	156
Studies	13	13	13	13

Notes: Panel A presents the results of funnel asymmetry test for the subset of estimates with GDP deflator as the measure of inflation and Panel B presents the results of the same test when CPI is used. WLS = weighted least squares. FE = study fixed effects. Study = estimates are weighted by the inverse of the number of estimates reported per study. Standard errors are clustered at the study level; 95% confidence intervals from wild bootstrap clustering are reported in square brackets, if applicable. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4.B2: Linear funnel asymmetry tests: labor share and output gap

Panel A: Labor share	WLS	FE	BE	Study
Standard error	-1.331^{***}	-0.051	-3.090^{***}	-1.872^{***}
(*publication bias*)	(0.462)	(0.865)	(1.004)	(0.606)
	[-2.485, -0.336]			[-3.167, -0.540]
Constant	0.837^{***}	0.783^{***}	0.890^{***}	0.851^{***}
(*mean beyond bias*)	(0.0178)	(0.037)	(0.032)	(0.028)
	[0.790, 0.883]			[0.769, 0.910]
Observations	403	403	403	403
Studies	29	29	29	29
Panel B: Output gap	WLS	FE	BE	Study
Standard error	-3.791^{***}	0.387	-3.905^{**}	-3.683^{***}
(*publication bias*)	(1.100)	(2.450)	(1.618)	(1.108)
	[-7.305, 0.014]			[-6.547, -0.099]
Constant	0.844^{***}	0.709^{***}	0.865^{***}	0.849^{***}
(*mean beyond bias*)	(0.022)	(0.079)	(0.036)	(0.021)
	[0.746, 0.911]			[0.780, 0.897]
Observations	45	45	45	45
Studies	12	12	12	12

Notes: Panel A presents the results of funnel asymmetry test for the subset of estimates when the forcing variable is labor share and Panel B presents the results when the output gap is the forcing variable. See Table 4.B1 for details.

Table 4.B3: Linear funnel asymmetry tests: GMM vs other estimators

Panel A: GMM	WLS	FE	BE	Study
Standard error	-1.807^{***}	0.441	-3.958^{***}	-2.186^{***}
(*publication bias*)	(0.576)	(1.062)	(1.184)	(0.607)
	[-3.399, -0.644]			[-3.438, -0.921]
Constant	0.848^{***}	0.754^{***}	0.915^{***}	0.849^{***}
(*mean beyond bias*)	(0.026)	(0.044)	(0.040)	(0.031)
	[0.793, 0.924]			[0.770, 0.915]
Observations	416	416	416	416
Studies	28	28	28	28
Panel B: Other estimators	WLS	FE	BE	Study
Standard error	-1.869^{**}	0.121	-2.802^{*}	-2.573^{**}
(*publication bias*)	(0.936)	(2.122)	(1.416)	(1.067)
	[-4.826, 0.209]			[-5.203, 0.116]
Constant	0.824^{***}	0.754^{***}	0.854^{***}	0.842^{***}
(*mean beyond bias*)	(0.044)	(0.074)	(0.032)	(0.025)
	[0.694, 0.921]			[0.761, 0.900]
Observations	93	93	93	93
Studies	15	15	15	15

Notes: Panel A presents the results of the formal funnel asymmetry test for the subset of parameters estimated by GMM and Panel B presents the results of the other estimators. See Table 4.B1 for details.

Table 4.B4: Linear funnel asymmetry tests: countries

Panel A: US	WLS	FE	BE	Study
Standard error	−0.967***	0.225	-2.146	−1.153**
(*publication bias*)	(0.370)	(0.878)	(1.401)	(0.535)
	[-2.034, -0.150]			[-2.239, -0.073]
Constant	0.798***	0.747***	0.840***	0.785***
(*mean beyond bias*)	(0.027)	(0.038)	(0.047)	(0.026)
	[0.695, 0.841]			[0.724, 0.846]
Implied duration (quarters)	4.950	3.953	6.250	4.651
Observations	303	303	303	303
Studies	25	25	25	25
Panel B: Europe	WLS	FE	BE	Study
Standard error	−3.186***	0.038	-3.895	−2.867***
(*publication bias*)	(0.700)	(3.807)	(2.234)	(0.873)
	[-6.767, -0.775]			[-6.354, -0.758]
Constant	0.891***	0.783***	0.894***	0.880***
(*mean beyond bias*)	(0.026)	(0.127)	(0.044)	(0.022)
	[0.791, 0.952]			[0.791, 0.952]
Implied duration (quarters)	9.174	4.608	9.434	8.333
Observations	93	93	93	93
Studies	10	10	10	10
Panel C: Oceania	WLS	FE	BE	Study
Standard error	2.121***	1.945	6.170	2.783***
(*publication bias*)	(0.514)	(0.881)	(2.620)	(0.959)
	[1.681, 6.653]			[1.376, 6.161]
Constant	0.780***	0.786***	0.599*	0.740***
(*mean beyond bias*)	(0.033)	(0.033)	(0.091)	(0.059)
	[0.614, 1.285]			[0.621, 0.876]
Implied duration (quarters)	4.545	4.673	2.494	3.846
Observations	48	48	48	48
Studies	3	3	3	3
Panel D: Asia	WLS	FE	BE	Study
Standard error	−4.311***	5.185*	−7.580*	−4.129***
(*publication bias*)	(1.625)	(1.921)	(2.760)	(1.509)
	[-16.140, 1.049]			[-13.510, 0.304]
Constant	0.773***	0.263*	0.956***	0.839***
(*mean beyond bias*)	(0.144)	(0.103)	(0.099)	(0.0727)
	[-37.460, 2.210]			[-2.581, 1.595]
Implied duration (quarters)	4.405	1.357	22.727	6.211
Observations	42	42	42	42
Studies	5	5	5	5

Notes: This table reports the results for different regions. See Table 4.B1 for details.

Table 4.B5: Linear funnel asymmetry tests: significant explanatory variables

Panel A: Hybrid NKPC	WLS	FE	BE	Study
Standard error	−1.844***	0.419	−2.933***	−1.986***
(*publication bias*)	(0.549)	(1.256)	(0.937)	(0.649)
	[-3.164, -0.634]			[-3.537, -0.581]
Constant	0.817***	0.722***	0.868***	0.834***
(*mean beyond bias*)	(0.028)	(0.053)	(0.028)	(0.024)
	[0.725, 0.875]			[0.764, 0.897]
Observations	284	284	284	284
Studies	27	27	27	27
Panel B: CB affiliation	WLS	FE	BE	Study
Standard error	−1.852**	-0.837	−3.680**	−2.718***
(*publication bias*)	(0.727)	(1.623)	(1.588)	(0.989)
	[-4.415, -0.332]			[-4.893, -0.635]
Constant	0.844***	0.805***	0.875***	0.848***
(*mean beyond bias*)	(0.035)	(0.062)	(0.039)	(0.027)
	[0.740, 0.922]			[0.752, 0.915]
Observations	221	221	221	221
Studies	21	21	21	21
Panel C: Open economy	WLS	FE	BE	Study
Standard error	-2.252	0.890	-2.619	-1.781
(*publication bias*)	(1.766)	(1.858)	(2.828)	(1.682)
	[-9.362, 1.774]			[-6.820, 1.920]
Constant	0.918***	0.814***	0.883***	0.857***
(*mean beyond bias*)	(0.074)	(0.061)	(0.073)	(0.044)
	[0.752, 1.15]			[0.746, 1.137]
Observations	84	84	84	84
Studies	8	8	8	8
Panel D: Inflation target	WLS	FE	BE	Study
Standard error	-1.126	1.094	−10.62***	-2.245
(*publication bias*)	(1.784)	(1.115)	(2.193)	(1.719)
	[-9.956, 2.098]			[-8.697, 2.142]
Constant	0.850***	0.759***	1.228***	0.834***
(*mean beyond bias*)	(0.050)	(0.046)	(0.093)	(0.078)
	[0.616, 1.188]			[0.501, 1.197]
Observations	82	82	82	82
Studies	8	8	8	8
Panel E: Model estimated	WLS	FE	BE	Study
Standard error	−7.688***	-0.278	−6.643**	−5.287***
(*publication bias*)	(1.340)	(4.547)	(2.148)	(1.307)
	[-12.470, -2.803]			[-9.087, -2.287]
Constant	0.921***	0.729***	0.904***	0.869***
(*mean beyond bias*)	(0.056)	(0.118)	(0.047)	(0.041)
	[0.657, 1.027]			[0.745, 0.996]
Observations	39	39	39	39
Studies	9	9	9	9

Notes: This table reports the results for different regions. See Table 4.B1 for details.

4.B.2 Nonlinear tests

Panel B of Table 1 in the paper and also Tables 4.B6-4.B9 present the results obtained from nonlinear techniques. Ioannidis et al. (2017) propose the Weighted Average of Adequately Powered (WAAP) technique, which considers the estimates when their statistical power is above an 80% threshold. In other words, by using the WAAP technique, we assign a weight to each estimate with adequate power to compute a weighted mean corrected for bias. Furthermore, Andrews and Kasy (2019) suggest the second nonlinear method used in this paper. This technique assumes that publication probability changes after crossing conventional t-statistic thresholds. This technique re-weights estimates in the vicinity of the threshold based on how they are present in the literature.

The Endogenous Kink (EK) technique proposed by Bom and Rachinger (2019), is the third nonlinear method used in the meta-analysis. This method extends the linear funnel asymmetry test by assuming that the selection of estimates for publication is constrained with particular precision cut-offs in each literature. Finally, Furukawa (2021) develops a stem-based method that considers only the most precise estimates (i.e., the stem of the funnel plot). The method considers both efficiency (increasing in the number of included estimates) and bias (decreasing in the number of included precise estimates) and optimizes the trade-off between them.

Table 4.B6: Nonlinear funnel asymmetry tests: GDP deflator and CPI

Panel A: GDP deflator	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.795***	0.792***	0.795***	0.848***
	(0.020)	(0.010)	(0.007)	(0.052)
Observations	353	353	353	353
Studies	28	28	28	28
Panel B: CPI	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.820***	0.749***	0.820***	0.820***
	(0.072)	(0.016)	(0.016)	(0.048)
Observations	156	156	156	156
Studies	13	13	13	13

Notes: This table reports the results of nonlinear techniques regarding different inflation measures. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4.B7: Nonlinear funnel asymmetry tests: labor share and output gap

Panel A: Labor share	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.805***	0.789***	0.805***	0.794***
	(0.018)	(0.009)	(0.006)	(0.051)
Observations	403	403	403	403
Studies	29	29	29	29
Panel B: Output gap	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.776***	0.639***	0.776***	0.770***
	(0.025)	(0.034)	(0.020)	(0.035)
Observations	45	45	45	45
Studies	12	12	12	12

Notes: This table reports the results of nonlinear techniques regarding different proxies of marginal costs. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4.B8: Nonlinear funnel asymmetry tests: GMM vs other estimators

Panel A: GMM	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.804***	0.792***	0.804***	0.823***
	(0.026)	(0.009)	(0.004)	(0.054)
Observations	416	416	416	416
Studies	28	28	28	28
Panel B: Others	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.789***	0.753***	0.789***	0.777***
	(0.042)	(0.023)	(0.014)	(0.032)
Observations	93	93	93	93
Studies	15	15	15	15

Notes: This table reports the results of nonlinear techniques regarding GMM and other estimators. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4.B9: Nonlinear funnel asymmetry tests: countries

Panel A: US	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.774***	0.756***	0.774***	0.791***
	(0.025)	(0.010)	(0.007)	(0.063)
Implied duration (quarters)	4.425	4.098	4.425	4.785
Observations	303	303	303	303
Studies	25	25	25	25
Panel B: Europe	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.832***	0.821***	0.831***	0.783***
	(0.035)	(0.019)	(0.016)	(0.036)
Implied duration (quarters)	5.952	5.587	5.917	4.608
Observations	93	93	93	93
Studies	10	10	10	10
Panel C: Oceania	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.826***	0.879***	0.826***	0.846***
	(0.028)	(0.011)	(0.014)	(0.047)
Implied duration (quarters)	5.747	8.264	5.747	6.494
Observations	48	48	48	48
Studies	3	3	3	3
Panel D: Asia	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.654**	0.446***	0.657***	0.588*
	(0.163)	(0.001)	(0.043)	(0.304)
Implied duration (quarters)	2.890	1.805	2.915	2.427
Observations	42	42	42	42
Studies	5	5	5	5

Notes: This table reports the results of nonlinear techniques for reported estimates based on different regions. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4.B10: Nonlinear funnel asymmetry tests: significant explanatory variables

Panel A: Hybrid NKPC	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.775***	0.764***	0.775***	0.813***
	(0.028)	(0.017)	(0.009)	(0.083)
Observations	284	284	284	284
Studies	27	27	27	27
Panel B: CB affiliation	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.800***	0.778***	0.800***	0.791***
	(0.029)	(0.012)	(.009)	(0.036)
Observations	221	221	221	221
Studies	21	21	21	21
Panel C: Open economy	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.865***	0.835***	0.865***	0.796***
	(0.054)	(0.017)	(0.154)	(0.065)
Observations	84	84	84	84
Studies	8	8	8	8
Panel D: Inflation targeting	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.823***	0.889***	0.823***	0.834***
	(0.044)	(0.022)	(0.189)	(0.144)
Observations	82	82	82	82
Studies	8	8	8	8
Panel E: Model estimated	Ioannidis et al. (2017)	Andrews and Kasy (2019)	Bom and Rachinger (2019)	Furukawa (2021)
Effect beyond bias	0.783***	0.640***	0.783***	0.809***
	(0.061)	(0.001)	(0.264)	(0.064)
Observations	39	39	39	39
Studies	9	9	9	9

Notes: This table reports the results of nonlinear techniques for subgroups of reported estimates based on the most decisive variables obtained from BMA. Standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.C Explanatory variables, summary statistics, and additional BMA results

Bayesian model averaging (BMA) method is a natural solution to model uncertainty within the Bayesian setting. Using all possible subsets of explanatory variables, BMA runs numerous regression models and forms a weighted average over all of them. If the set of explanatory variables contains n variables, there will be combinations of 2^n variables and 2^n models. Defining $\mathcal{P}(M_i)$, $\mathcal{P}\left(y \mid M_i, X_i\right)$, and $\mathcal{P}(y \mid X_i)$ as the model prior, the marginal likelihood, and the integrated likelihood, respectively, posterior model probabilities (PMP) are obtained as follows:

$$\mathcal{P}\left(M_i \mid y, X\right) = \frac{\mathcal{P}\left(y \mid M_i, X\right)\mathcal{P}\left(M_i\right)}{\mathcal{P}(y \mid X_n)} \equiv \frac{\mathcal{P}\left(y \mid M_i, X\right)\mathcal{P}\left(M_i\right)}{\sum_{s=1}^{2N}\mathcal{P}\left(y \mid M_s, X_s\right)\mathcal{P}\left(M_s\right)},$$

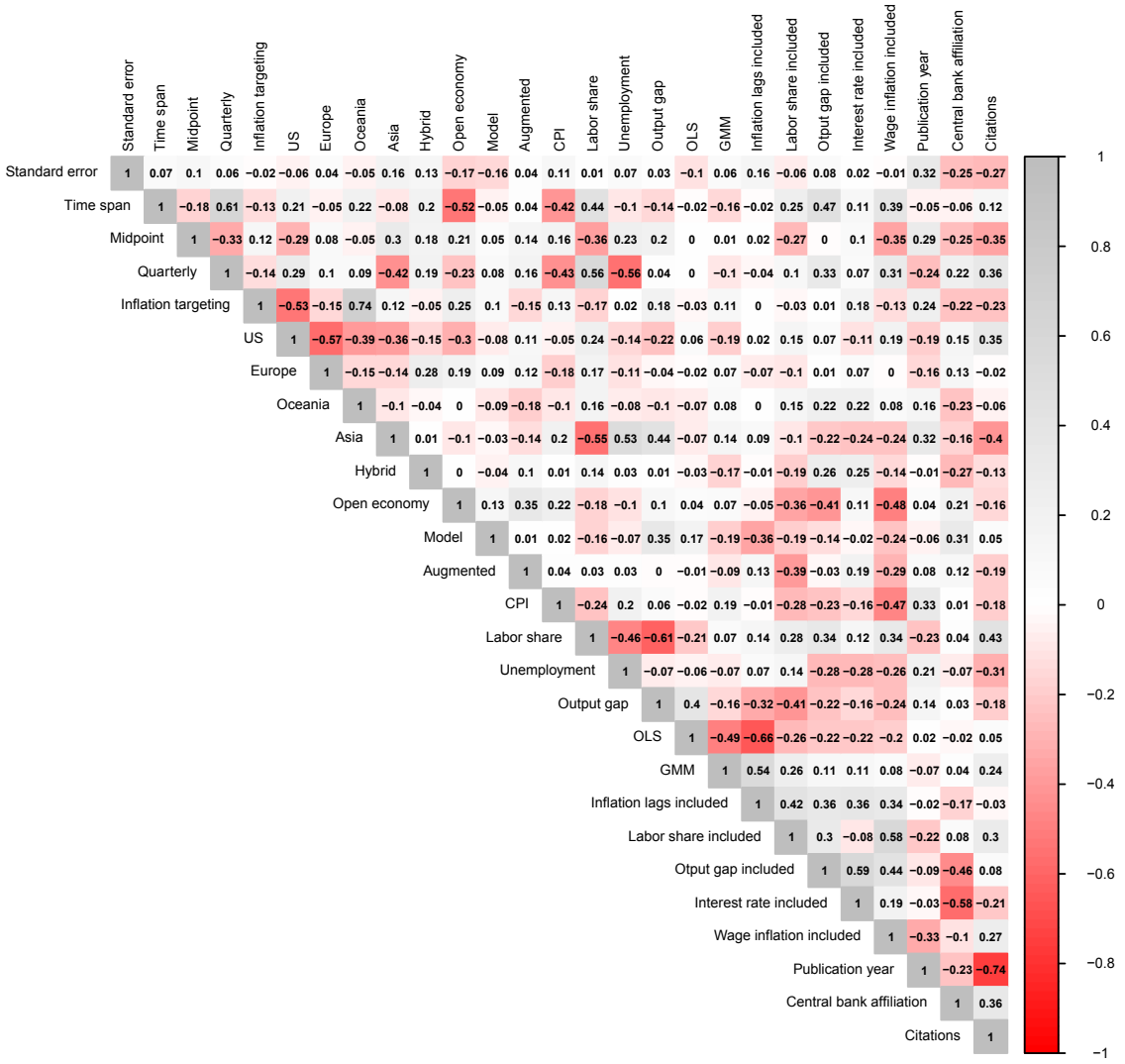
The model weighted posterior distribution for θ can be written as:

$$\mathcal{P}(\theta \mid y, X) = \sum_{i=1}^{2^n}\mathcal{P}\left(\theta \mid M_i, y, X\right)\mathcal{P}\left(M_i \mid y, X\right).$$

The model prior is a key factor in conducting BMA since it reflects the prior beliefs about the model. The benchmark prior is dilution prior suggested by George (2010), which takes into account the collinearity of variables in each model by assigning higher weights to models that exhibit lower collinearity. Additionally, for robustness checks, BRIC g-prior and HQ g-prior are used. The former is the benchmark g-prior for parameters with the beta-binomial model prior, while the latter asymptotically mimics the Hannan-Quinn criterion. For more information, Steel (2020) provides a comprehensive and insightful summary of model averaging in economics.

4.C.1 Explanatory variables

Figure 4.C1: Correlation matrix



Notes: The figure shows Pearson correlation coefficients for the explanatory variables described in Table 4.C1.

Data characteristics. I control for the time span in which the Calvo parameter is estimated. I also control for the frequency of data by including dummy variables indicating whether quarterly data is used. There are dummy variables reflecting the region of the data source used in the estimation: the US, Europe, Oceania, and Asia.

Specifications. Controlling for model specifications, I codify a dummy variable capturing if the estimate is obtained within the hybrid NKPC or a purely forward-looking NKPC setting. Besides, two other dummy variables indicate if the reported estimate is obtained within an open economy setting or an augmented NKPC setting (i.e., the NKPC includes other terms in addition to expected inflation and economic activity). I also codify a dummy variable reflecting if the Calvo parameter is estimated within a model. Estimating the NKPC and, in particular, the Calvo parameter is sensitive to the choice of inflation measurement; see, e.g., (Mavroeidis et al. 2014). Hence, I introduce a dummy variable accounting for CPI and GDP deflator as inflation measurements. As discussed by Galí and Gertler (1999), the choice of a valid proxy for marginal cost can affect the estimated parameters within the NKPC. Three dummy variables control for marginal costs proxies: labor share, unemployment, and the output gap.

Estimation techniques. There are seven dummy variables defined to capture different aspects of estimation methods. Two dummy variables denote the ordinary least squares (OLS) and the generalized method of moments (GMM) methods used in estimating the parameter, which are used for 87% of the estimates in the sample. Moreover, I include five dummy variables reflecting the instruments used in estimating the parameter.

Table 4.C1: Definition and summary statistics of explanatory variables

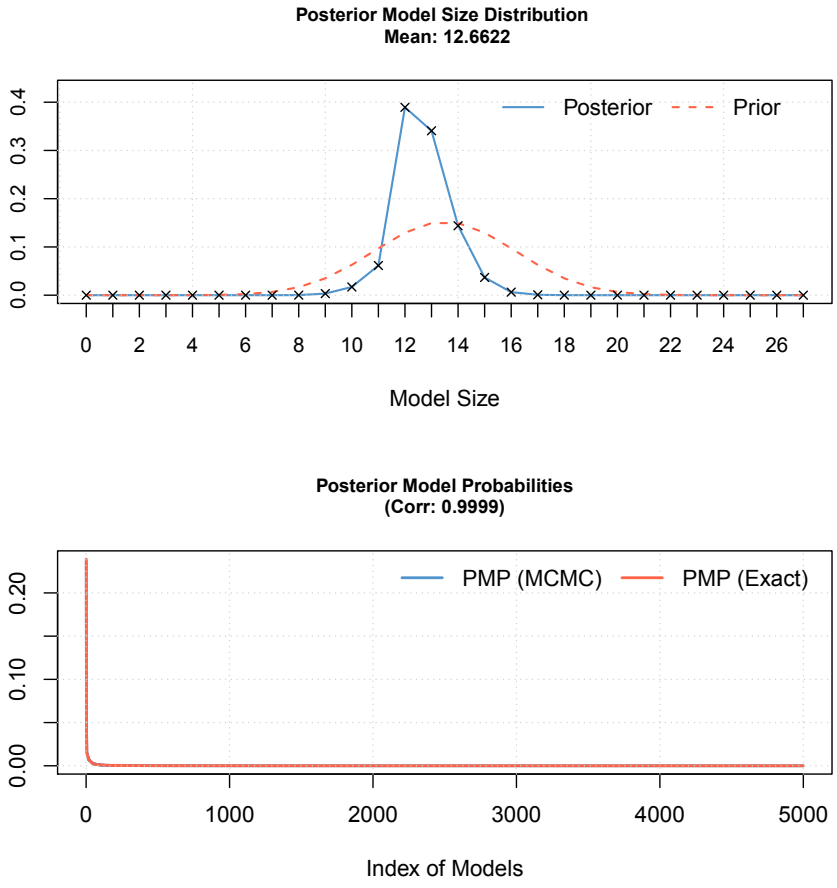
Variable	Description	Mean	SD	No. papers
θ	The estimated Calvo parameter in the NKPC equation.	0.72	0.25	-
Standard error	The standard error of the estimated coefficient of inflation expectations.	0.31	2.69	-
Data characteristics				
Time span	The logarithm of the data time span used to estimate θ.	3.37	0.52	-
Midpoint	The logarithm of the median year of the data used minus the earliest median year in primary studies.	2.80	0.75	-
Quarterly	= 1 if the data frequency is annual (reference category: monthly/annual).	0.92	0.26	37
Inflation targeting	=1 if the central bank employs an inflation targeting regime during at least half of the estimation period.	0.16	0.37	8
US	= 1 if the estimate is for the U.S. (reference category: other countries).	0.99	0.49	25
Europe	= 1 if the estimate is for European countries (reference category: other countries).	0.18	0.39	10
Oceania	= 1 if the estimate is for Australia and New Zealand countries (reference category: other countries).	0.10	0.29	3
Asia	= 1 if the estimate is for Asian countries (reference category: other countries).	0.08	0.27	5
Specifications				
Hybrid	= 1 if the estimate is from a hybrid NKPC setting (reference category: purely forward-looking NKPC).	0.56	0.50	27
Open economy	= 1 if the estimate is from an open economy specification (reference category: closed economy).	0.16	0.37	8
Model	= 1 if θ is estimated within a model.	0.08	0.27	9
Augmented	= 1 if the NKPC includes other terms in addition to expected inflation and economic activity.	0.24	0.43	10
CPI	= 1 if CPI is the measure of inflation (reference category: GDP deflator).	0.31	0.46	13
Labor share	= 1 if the labor income share (unit labor costs) is a proxy for marginal costs (reference category: other proxies).	0.79	0.41	29
Unemployment gap	= 1 if unemployment is a proxy for marginal costs (reference category: other proxies).	0.05	0.22	2
Output gap	= 1 if output gap is a proxy for marginal costs (reference category: other proxies).	0.09	0.28	12
Estimation techniques				
OLS	= 1 if the ordinary least square (OLS) method is used for the estimation (reference category: other methods).	0.05	0.22	8
GMM	= 1 if the generalized method of moments (GMM) is used for the estimation (reference category: other methods).	0.82	0.39	28
inflation lags included	= 1 if inflation lags are among instruments (reference category: inflation lags not among instruments).	0.91	0.28	30
Labor share included	= 1 if labor income share is among instruments (reference category: labor share not among instruments).	0.65	0.48	20
Output gap included	= 1 if the output gap is among instruments (reference category: Output gap not among instruments).	0.57	0.49	20
Interest rate included	= 1 if the interest rate is among instruments (reference category: interest rate not among instruments).	0.58	0.49	18
Wage inflation included	= 1 if wage inflation is among instruments (reference category: Wage inflation not among instruments).	0.54	0.50	16
Publication characteristics				
Publication year	The logarithm of the publication year of the study minus the publication year of the first primary study.	2.32	0.68	-
Central bank affiliation	= 1 if at least one of the authors is affiliated with a central bank.	0.43	0.50	21
Citations	The logarithm of the number of per-year citations of the study, according to Google Scholar.	1.40	1.51	-

Notes: SD = standard deviation No. papers = the number of papers that capture the dummy variable. The table excludes the definition and summary statistics of the reference categories, which are omitted from the regressions.

Publication characteristics. There is a variable for the publication year to capture the fact that a recent study is more likely to provide more accurate results since it employs newer theoretical and empirical methods. As a proxy accounting for the ex-post quality of the study, there is an explanatory variable denoting the number of citations of each study. Finally, I codify a dummy variable indicating if at least one of the authors is affiliated with a central bank. This variable helps capture possible workplace bias.

4.C.2 Robustness checks

Figure 4.C2: Model size and convergence for the benchmark BMA model



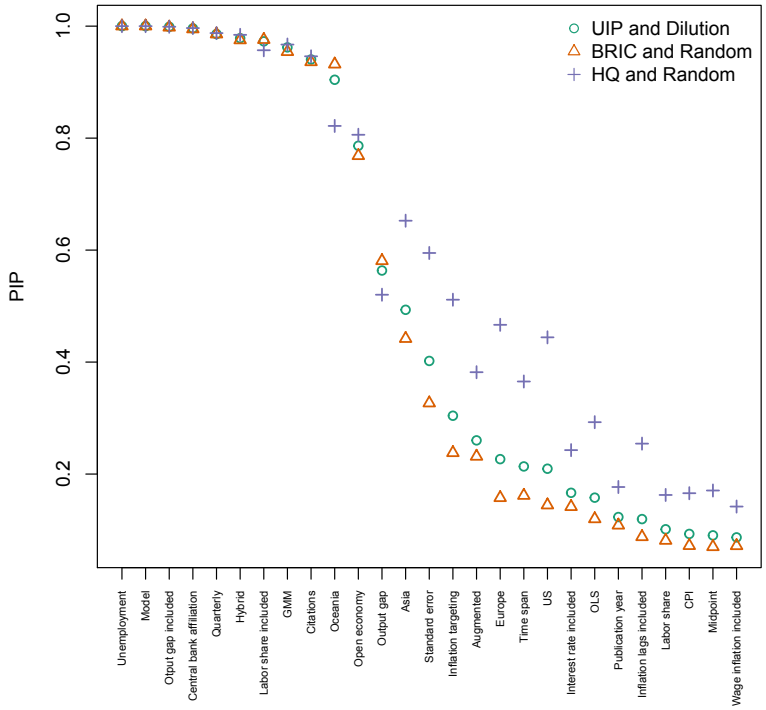
Notes: The figure illustrates the posterior model size distribution and the posterior model probabilities of the weighted BMA exercise reported in Table 2.

Table 4.C2: Alternative BMA priors

	BRIC g-prior			HQ g-prior		
Variable	Post. Mean	Post. SD	PIP	Post. Mean	Post. SD	PIP
Constant	0.959	N.A.	1.000	0.957	N.A.	1.000
Standard error	0.074	0.121	0.327	0.090	0.129	0.394
Data characteristics						
Time span	0.007	0.019	0.162	0.009	0.021	0.200
Midpoint	-0.001	0.004	0.070	-0.001	0.004	0.077
Quarterly	-0.198	0.061	0.986	-0.200	0.060	0.989
Inflation targeting	-0.018	0.038	0.238	-0.022	0.041	0.285
US	-0.020	0.063	0.145	-0.026	0.070	0.176
Europe	-0.022	0.064	0.158	-0.028	0.071	0.189
Oceania	0.174	0.061	0.932	0.171	0.065	0.920
Asia	-0.062	0.087	0.442	-0.067	0.090	0.471
Specifications						
Hybrid	-0.067	0.020	0.975	-0.068	0.019	0.989
Open economy	0.066	0.044	0.768	0.069	0.042	0.806
Model	-0.175	0.032	1.000	-0.175	0.032	1.000
Augmented	0.013	0.027	0.232	0.012	0.026	0.238
CPI	-0.001	0.007	0.072	-0.001	0.007	0.074
Labor share	-0.001	0.015	0.081	-0.001	0.015	0.083
Unemployment	-0.366	0.065	1.000	-0.364	0.062	1.000
Output gap	-0.056	0.055	0.581	-0.055	0.054	0.577
Estimation techniques						
OLS	-0.008	0.028	0.120	-0.010	0.031	0.142
GMM	-0.093	0.032	0.954	-0.095	0.030	0.974
Inflation lags included	-0.006	0.026	0.088	-0.006	0.027	0.093
Labor share included	-0.104	0.028	0.976	-0.103	0.027	0.980
Otput gap included	0.124	0.027	0.998	0.125	0.026	1.000
Interest rate included	0.007	0.020	0.142	0.007	0.020	0.153
Wage inflation included	0.002	0.008	0.072	0.002	0.008	0.074
Publication characteristics						
Publication year	-0.003	0.011	0.109	-0.002	0.010	0.100
Central bank affiliation	0.095	0.025	0.995	0.095	0.025	0.998
Citations	0.027	0.010	0.936	0.027	0.009	0.954
Observations	509			509		
Studies	40			40		

Notes: The response variable is the estimated Calvo parameter. SD = standard deviation, PIP = Posterior inclusion probability. The left panel applies BMA based on BRIC g-prior (the benchmark g-prior for parameters with the beta-binomial model prior). The right panel reports the results of BMA based on HQ g-prior, which asymptotically mimics the Hannan-Quinn criterion. Table 4.C1 presents a detailed description of all the variables.

Figure 4.C3: Posterior inclusion probabilities across different prior settings



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. The HQ prior asymptotically mimics the Hannan-Quinn criterion.

Chapter 5

Conclusion

This dissertation has embarked on a comprehensive exploration of sources of variation in estimating deep parameters in economics through the lens of meta-analysis. This powerful tool synthesizes findings from multiple studies to extract underlying trends and biases in economic research. As emphasized in the Lucas critique, deep parameters are foundational in economic modeling, offering insights into the stable and invariant aspects of individual behavior and economic processes crucial for policy modeling and forecasting. This work focuses on three distinct structural parameters, relative risk aversion, Frisch elasticity of labor supply, and the Calvo parameter, and it provides a broad spectrum of analysis on how data characteristics and other aspects of methodology and research design affect estimating such parameters. More specifically, this dissertation sheds light on challenges such as publication bias and estimation heterogeneity accompanying their empirical determination.

The first article provides a meta-analysis on relative risk aversion and highlights a significant divergence between the estimates and calibrations used in economic versus finance contexts, suggesting different risk tolerance levels perceived or utilized across these fields. Publication bias and the methodological variations of the studies significantly influence this divergence. The corrected

mean estimates after accounting for such biases suggest a larger estimate in the finance context.

The second article examines the Frisch elasticity of labor supply at intensive and extensive margins. This article uncovers the dual challenge of publication and identification biases due to methodological differences across studies. The findings emphasize the need for cautious calibration of economic models, advocating for reliance on quasi-experimental evidence and well-identified primary studies over simplistic averages of reported estimates.

The third article's analysis of the Calvo parameter through Bayesian model averaging reveals how the choice of model specifications and research design impact the estimation of this parameter within the empirical New Keynesian Philips curve. The insights derived from the linear and nonlinear techniques and the BMA method highlight the importance of context-specific calibration of the Calvo parameter to avoid biases toward conventionally used values.

In conclusion, by integrating advanced statistical techniques like Bayesian model averaging with meta-analysis, this dissertation addresses publication bias and considers research design characteristics that impact the estimation of deep parameters. Combining findings across multiple studies offers a subtle understanding of parameters essential for macroeconomic modeling and policy-making. Hence, this dissertation contributes to meta-analysis and macroeconomics literature. However, despite the significant development in understanding and correcting for biases in estimated deep parameters, this dissertation does not fully address crucial areas such as p-hacking and attenuation bias. These topics present avenues for future research.

Appendix A

Response to Opponents

I thank the reviewers and the committee for their insightful comments on the pre-defense version of my dissertation. The comments are typeset in *italics*, while my response is in Roman type. To maintain brevity, I have selectively included only those excerpts from the reports that require a response or necessitate revisions. Since the first chapter is now the Introduction, tables and figures numbers are updated (e.g., Table 1.2 to Table 2.2).

Response to Opponent: Professor Tom Stanley (Deakin University)

I am deeply grateful for your thorough review and the encouraging remarks on my dissertation thesis. Your assessment not only affirms the direction of my research but also inspires continued exploration and rigor in my work. I am looking forward to incorporating your valuable feedback into my final revisions.

Response to Opponent: Professor Heiko Rachinger (University of the Balearic Islands)

Article 1:

1. *In the abstract, I miss a reference to "meta-analysis or meta-regression".*
   Thank you for this comment. We will modify the abstract with an emphasis on meta-analysis.

2. *The introduction arrives in my opinion too fast to the meta-analysis. It might be helpful to start with a short discussion on why measuring the risk aversion matters and for what these estimates are used in the macro literature, beyond stating "Risk aversion is a key concept in economics and finance". You might want to copy something from "is important since they affect decisions on savings/investing and, consequently, asset prices in the economy. For instance, the degree of risk aversion plays a crucial role in the capital asset pricing model (CAPM) or consumption capital asset pricing model (CCAPM) since it heavily affects the investor's consumption and wealth portfolio, which ultimately alter asset prices." from Appendix 1B to the main text.*
   This is a valid point. We will add more details on the importance of RRA in the article's introduction before submission to a journal.

3. *I also feel that in the first paragraph some references could help.*
   We will revise the first paragraph and add a set of references from the current literature regarding the estimation of RRA and relevant empirical and theoretical issues.

4. *One of the recurrent findings is the difference between estimates of the relative risk aversion in the economics and the finance literature. An ad-*

ditional discussion of the very different estimates and calibration values in both literatures might be helpful. Where do these differences come from? Is there some rationale?

This is again a good point. Although finding the true sources of divergence between finance and economic estimates is challenging, we will incorporate more related details into the introduction. We argue that the possible sources of divergence between the two fields can be: a) the difference between modeling context in each field (e.g., finance literature is more of partial equilibrium models, while in economics, researchers mainly deal with the general equilibrium), and b) finance literature deals with investors, while economics deal with a more diverse population.

5. *When describing the data collection procedure you might want to refer to the guidelines (Havránek et al. 2020 and Irsova et al. 2024)*

   Sure, we will cite the relevant studies.

6. *Table 1.2: is it surprising that with the FE, differences between the economics and finance literature seem to disappear?*

   It is not particularly surprising that differences between the economics and finance literature disappear when using FE as FE models control for all time-invariant characteristics within each field. At the same time, BE and WLS are sensitive to other aspects of data, such as permanent differences between the fields or sample size and variance. However, even in this specification, the mean estimate in finance is still marginally larger than in economics. Hence, our conclusion based on this table still holds.

7. *Could you elaborate on why explicitly allowing for heterogeneity among the estimates relaxes the uncorrelation assumption, adding"insofar as..." after the first sentence in Section 1.4.*

   The more unambiguous statement would be: apart from considering only

one source of variation among estimates, by allowing for heterogeneity, one can investigate the variations in effect sizes from different studies that are due to various study populations, methodologies, or other factors.

8. *Why are the* 95% *credible intervals so wide? How confident are you then for the found effects and differences (between economics and finance for example)?*

   This is a good point. One of the limitations of results regarding the implied risk aversion is the wide credible intervals. It partially reflects the wide variation in the reported estimates. I will include more explanation in the final version of the paper to ensure that the reader considers the underlying uncertainty in the results.

9. *Could you explain a little more the values in Table 1.6.*

   The table reports the subjective values of relative risk aversion in the literature if there was no publication bias in addition to homogeneity in the context in which they are estimated. We compute fitted values of the Frisch elasticity by plugging in our subjective preferred weights of collected explanatory variables. By doing so, we consider the characteristics of the research framework based on collected explanatory variables. For the overall best practice, for example, 3.73 is our subjective value suggested for calibration when no distinctive population type is considered in macroeconomic modeling. This value is lower than more common values used in the literature (e.g., 5). The rest of the table reports the values based on the same weights but for different population subsets, data frequency, or preferred method. We will incorporate additional explanations into the final version of the paper.

10. *Doesn't the conclusion "Table 1.6 shows that such an exercise yields imprecise results, but the point estimate for economics is still around 1,*

consistent with our previous results. The implied estimate for finance is somewhat larger, around 7, but not far from the 2–6 range discussed in the previous section." somehow not correspond to the quite smaller estimate using p-uniform which was your preferred estimate. There is a similar statement in the conclusion.*

You are right. I should formulate this statement more clearly. However, an explanation for having larger values in Table 2.6 is that even though p-uniform* relaxes the uncorrelation assumption, it still does not explicitly consider other aspects of heterogeneity. Addressing other research characteristics directly would give us the results reported in Table 2.6 that are essentially different from those solely based on p-uniform*.

11. *Additional robustness checks, especially regarding the distinction between economics and finance could be helpful. Are there studies in which this distinction is more straightforward than in others?*

    I acknowledge that our classification of studies into economics and finance fields is crude. However, the significant difference between the estimates between the two fields confirms our prior regarding the classification. As you suggested, we can conduct additional robustness checks, such as a categorization solely based on methodology. We will incorporate the additional exercises into the article's final version before submission.

12. *The tests of p-hacking in Table 1.C1 do not reject for Economics and Finance separately, but do so for all studies together. Does Elliott et al. (2022) give some rule of thumb on the necessary sample size?*

    This is again a valid point. Unfortunately, I face some difficulties utilizing the replication package provided by Elliott et al. (2022) due to a lack of instructions, including the rule of thumb on the sample size. Hence, the

results might be subject to some errors. However, co-authors and I will rerun the provided code before submission to detect potential errors.

13. *Table 1.C3: Why would you expect such a huge positive relationship between effect sizes and standard errors in cases when publication selection is no issue?*

    It is reasonable to consider that specific aspects of data or research framework could influence both estimates and standard errors, thus producing a correlation even without the presence of publication bias. Therefore, this is a motivation to investigate the effect of other aspects of data and methods impacting the reported estimates, leading us to the Heterogeneity section of the article. We state the same argument in the first paragraph of Page 23 in the new version of the dissertation.

14. *Is any of the information in Table 1.D1 especially informative?*

    We provide this table to summarize the BMA practice for the reader interested in the computational aspects of the procedure.

Article 2:

1. *Are you sure about "Since the t-statistic is a ratio of the point estimate to the corresponding standard error and since the symmetry property implies that the numerator and denominator are statistically independent quantities, it follows that estimates and standard errors should not be correlated."? I guess symmetry implies linear uncorrelatedness, but not necessarily independence.*

    Yes, you are right. The linear uncorrelation does not necessarily imply independence. We should formulate the statement more precisely.

2. *The statement "This does not help alleviate publication bias since working*

papers are intended for eventual publication and any mechanisms that lead to preferences for positive or significant estimates in journal articles also apply to working papers, as shown, for example, by Rusnak et al. (2013)" seems too strong. Some mechanisms definitely apply to working papers, but all of them?

Again, you are right. The statement is a strong claim. There are indeed exceptions. However, in this meta-analysis, the statement can be almost correct as working papers are written by researchers who are actively publishing. Hence, we assume they intend to publish the working papers.

3. *"Using the outliers at their face value or omitting them from the analysis does not change our main results qualitatively". Is too vague a statement. In which sense?*

   The primary purpose is to state that our main results regarding the publication bias size are consistent whether we use the original, winsorized, or truncated (censored) dataset.

4. *You state that "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." Both effects seem to go in different directions. Can they be somehow disentangled?*

   This is a good point. For publication bias, we can conduct a series of tests on the subsets of estimates sorted by citations and quality of journals (i.e., top five vs others). Regarding BMA, an interaction term between standard error and *Top journal* might help disentangle the effects.

5. *In Section 2.4.2, the word "incorrect" is too strong in "regress the reported elasticities on all the variables introduced above. But that approach is incorrect because it ignores model uncertainty".*

Yes, we should be more cautious in writing this claim and would use "not always necessarily correct" instead.

6. *The finding that "either IV or non-parametric techniques used in estimating the elasticity affect the results systematically" could be more stressed.* Sure! We could expand this finding more in comparison with the other decisive variable in the same category, "Quasi-experimental". However, our main focus is on the decisive variables with a PIP higher than 0.75.

7. *For the 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." Is this really conservative?*
   This is a good question. Indeed, the definition of conservative in selecting weights for a subjective measure can be subjective itself. In addition to publication bias, we are mainly concerned with other biases for the overall elasticity. That is why we give a higher weight to the IV estimates. Regarding other variables, we base our assumptions on evidence pervasive in the literature to remain as neutral as possible.

8. *For Table 2.6, how are the confidence intervals obtained? Taking the uncertainty regarding the estimates into account but not the one regarding the best practice values? Discuss it a little more. Aren't then all values not distinguishable from 0?*
   The 95% confidence intervals are computed using `lincom` command in `Stata`, where we consider the results of frequentist model averaging, as this method naturally deals with uncertainty by testing several models with all possible combinations of included parameters. In the case of extensive margin, most values are around zero. However, the values at the intensive margin are clearly distinguishable from zero, which means the aggregate elasticity is significantly larger than zero.

9. *On page 132, I would not use the notion "spurious precision" when talking about misleading precision, when recommending 0.2 for the calibration.*
   Sure, using *"spurious precision"* can be interpreted wrongly. I will revise the wording in the final version of the article.

10. *Are the confidence intervals of the MAIVE in Table 2.A2 correct? They do not include the estimate?*
    Thank you for this comment. There are indeed typos in the confidence intervals of the MAIVE. We will correct and report them to the journal.

11. *Is the OLS frequentist check in Table 2.A5 really a robustness check given that the same variables as the ones chosen in the BMA are used?*
    The rationale is to study whether or not the decisive variables found in BMA are significant in a separate frequentist approach. So, one can interpret this practice as a robustness check for those variables' importance in determining the elasticity's size.

Article 3:

1. *Here the direction of the publication bias seems to be not so clear. Both Funnel plot and the tests in Table 3.2 suggest that the true effect is larger than the observed mean. However, all estimates in the region above 1 would be significant anyhow. So is it really the relationship between effect and standard error that drives the publication bias? In other words, why would it be the imprecise estimates that are dropped?*
   This is a valid point. The interpretation of results could be more explicit here. One practice could be testing subsets for significant and insignificant estimates to help understand if the relationship between effect and standard error is the primary driver of publication bias in this case.

2. *It would be nice if results such as "Unemployment and output gaps are associated with lower values of the Calvo parameter" in Section 3.4 were further discussed.*

   Yes, you are right. However, because of the journal's format and word limit, it was not possible to provide more details about the results.

3. *In the conclusion, specifying what are the "more orthodox values of the Calvo parameter" and which is the "microeconomic data evidence" would contribute to the readability.*

   As mentioned in the first paragraph of the introduction, an orthodox value of the Calvo parameter is 0.75. It would be clearer if I provided a more explicit value in the conclusion, too. Again, I have provided some references regarding the microeconomic evidence in the introduction, which could be restated in the conclusion

4. *A limitation such as "Further studies can extend the framework of this paper by investigating other aspects of research design to estimate the Calvo parameter absent from this paper." raises the question of why it has not been done here yet.*

   You are right again. Actually, there is limited literature studying the precision of the Calvo parameter (see, e.g., Dufour et al. 2010). However, to my knowledge, these studies either focus on a purely theoretical framework or empirical evidence from specific regions, making them different from this paper's approach.

I sincerely thank you for reading my thesis and for your positive comments on its content. Your detailed report and insightful feedback are greatly appreciated. I will incorporate all your comments into the first article before submitting it for publication.

Response to Opponent: Professor Patrice Laroche (Université de Lorraine)

Article 1:

- *I was wondering to what extent stopping the search for primary studies for reasons of feasibility would not lead to studies that could have been part of the study sample being left aside (in the drawer). It is very difficult to inspect 3,500 studies, but are Google Scholar's algorithms sufficiently reliable to justify this methodological choice? Finally, did the author combine this bibliographic search in Google Scholar with other bibliographic databases and other traditional methods of searching for articles?*

  We choose Google Scholar because of its universal coverage and full-text capabilities. Based on our experience, stopping the inspection of studies after the 1500^{th} study is reasonable, as we obtained utterly irrelevant results for the last hundreds of studies. Moreover, after sorting the results by citations and publication date, we compare our results from Google Scholar with those from the Web of Science. The comparison gives us an almost identical outcome.

- *Prior research has mentioned the paucity of information on interrater reliability (IRR) including the number of coders involved, at what stages and how IRR tests were conducted, and how disagreements were resolved. Findings indicated that coding behavior changes both between and within individuals over time, emphasizing the importance of conducting regular and systematic IRR and interrater reliability tests, especially when multiple coders are involved, to ensure consistency and clarity at the screening and coding stages. So I wondered whether the candidate could provide more information on this subject.*

Although we cannot provide exact statistics regarding IRR, we did our best to decrease errors during the search and coding procedures as much as possible. To reduce the danger of mistakes, two co-authors collect the data independently, and the third co-author resolves inconsistencies between these two datasets. Thank you for mentioning this critical issue. I will indeed consider it for future research.

- *The exploration of heterogeneity by adopting a BMA is flawless in this study. My main comment is that the estimates may come from the same study and I wonder how the BMA procedure controls for estimation independence bias in this case. It is a question I have always wondered about and I do not have the answer.*

  You are right. One of the problems with BMA in this context is the bias for the estimates within studies. Unfortunately, to my knowledge, there is no method or prior considering this issue. One robustness check could be conducting BMA that includes only the median estimate of each "*estimation method-study*". The (in)consistency in the results might be helpful in understanding the extent of estimation independence bias.

- *The results of OLS (Table 1.5, page 24) used for verification are clustered at the study level, but is this sufficient? Would it be possible to propose other estimates as robustness tests to further control this type of bias?*

  Clustering at the author level can be another control for bias. However, in our dataset, except for two papers (Hyde and Sherif 2005a and Hyde and Sherif 2005b), even in some studies with the same authors, the methods used are different from each other. Hence, the clustering at the author level should provide similar results.

Article 2:

- *My initial remarks [regarding the first article] could just as easily be repeated for this second contribution.*

  Please see the responses to your comments regarding the first article.

Article 3:

- *In this article, as in the others, we could look more closely at publication bias and, in particular, the issue of p-hacking. There is a whole literature on p-hacking that could have been used to go further (Brodeur et al. 2020; 2023). That said, this dimension could be the subject of future research.*

  Thank you for your useful comment. As you noticed, a limitation in this article and my dissertation in general is investigating *p*-hacking. I address *p*-hacking in a limited manner in this dissertation (e.g., using the method proposed by Elliott et al. 2022 in Table 2.C1). However, for future research, I plan to address both publication bias and *p*-hacking by using the methods developed in the literature you mentioned.

Overall comments:

- *It might be interesting to have a presentation (in the form of robustness tests) of meta-analytical results using the Stanley and Doucouliagos approach, for example (FEE-WLS; Random Effects, etc.).*

  Thank you for your comment. Although I do not report the RE results in this dissertation, the results in articles 1 and 3 (e.g., Tables 2.2 and 4.2) are basically WLS and FE-WLS proposed by Stanley and Doucouliagos.

- *I would recommend that the candidate write an introduction and a general conclusion in order to show the interest of Bayesian approaches compared to others at the heart of his work as a meta-analyst and in order to explain*

the common thread that links each of his empirical studies.

The updated dissertation includes a general introduction and conclusion.

I greatly appreciate your time reading my thesis and your kind remarks about its content. Thank you for your detailed report and valuable feedback. Before submitting it to a journal, I will incorporate your insightful comments into the unpublished article.

References

Brodeur, A., Carrell, S., Figlio, D., and Lusher, L. (2023). Unpacking p-hacking and publication bias. *American Economic Review*, forthcoming.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and causal inference in economics. *American Economic Review*, 110(11):3634–3660.

Dufour, J.-M., Khalaf, L., and Kichian, M. (2010). On the precision of Calvo parameter estimates in structural NKPC models. *Journal of Economic Dynamics and Control*, 34(9):1582–1595.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). Detecting p-hacking. *Econometrica*, 90(2):887–906.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K., and Van Aert, R. (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys*, 34(3):469–475.

Hyde, S. and Sherif, M. (2005a). Consumption asset pricing models: Evidence from the UK. *The Manchester School*, 73(3):343–363.

Hyde, S. and Sherif, M. (2005b). Don't break the habit: structural stability tests of consumption asset pricing models in the UK. *Applied Economics Letters*, 12(5):289–296.

Irsova, Z., Doucouliagos, H., Havranek, T., and Stanley, T. (2024). Meta-Analysis of Social Science Research: A Practitioner's Guide. *Journal of Economic Surveys*, (forthcoming).

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