Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

Dissertation Thesis

Three Essays on the Economics of Education

Author: **Mgr. Olesia Zeynalova**

Advisor: **doc. PhDr. Zuzana Havránková Ph.D**

Academic Year: **2023**

Declaration of Authorship

The author hereby declares that she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, March 31, 2023 Olesia Zeynalova

Abstract

This dissertation focuses on three aspects of higher education policy that are pertinent to both deans and policymakers. The first essay contributes to the debate about the relationship between tuition fees and demand for higher education using meta-analysis. While large negative estimates dominate the literature, we show that researchers report positive and insignificant estimates less often than they should. After correcting for this publication bias, we find that the literature is consistent with the mean tuition-enrollment elasticity being close to zero. The second essay shows how easily firms can substitute workers with different educational backgrounds. We evaluate the elasticity of substitution between skilled and unskilled workers, which is a key parameter in the analysis of wage inequality. We show that the empirical literature is consistent with both publication and attenuation bias in the estimated inverse elasticities. The publication bias-corrected estimates remain close to zero. The result is consistent with attenuation bias in the literature and implies an elasticity of 4 after correction for both biases. The third essay shows how the real implementation of the Russian Unified State Exam (USE) reform for high school graduates affected the returns to university education. The findings suggest that the reform has positively impacted education returns and it is more prominent in Moscow and St. Petersburg, where most elite schools are located. This could be due to the increased mobility of talented individuals from small cities, towns, and rural areas to bigger cities, where salaries are higher.

Keywords: Demand for higher education, enrollment, tuition fees, elasticity of substitution, skill premium, unified state exam, return on education

JEL Codes: I23, I26, I28, J23, J24, J31, J30

Acknowledgment

I would like to express gratitude to my supervisor, doc. PhDr. Zuzana Havrankova Ph.D., who guided me throughout this research project and also co-authored two of my research articles. I appreciate all the valuable comments, support and encouragement. This dissertation would not be possible without the hard work and contribution of my other dear co-authors. Therefore, special thanks goes to prof. PhDr. Tomas Havranek Ph.D. and Lubica Laslopova. Last but not least, I am grateful to my family, especially my parents, spouse and children, for their patience, inspiration and moral support throughout years of study.

Typeset in $\LaTeX\,2_\varepsilon$ using the IES Thesis Template.

Bibliographic Record

Zeynalova, Olesia: *Three Essays on the Economics of Education.* Dissertation thesis. Charles University in Prague, Faculty of Social Sciences, Institute of Economic Studies. March 2023, pages 179.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

According to a recent report by the World Bank (2018), human capital (measured in terms of income) comprises 64 percent of the total global wealth. Education constitutes a major factor in enhancing the quality of human capital (as noted by Hanushek & Woessmann 2012). This dissertation focuses on three aspects of higher education policy that are pertinent to both deans and policymakers. The first essay shows how much students are discouraged by payments for university education, the second essay shows how easily firms can substitute workers with different educational backgrounds, and the third essay shows how the real implementation of a unified final exam for high school graduates affected the returns to university education. In all cases, policies aimed at promoting the quality and efficiency of education, and thus enhancing the quality of human capital, yield unintended consequences in different contexts.

Chapter 2 of this dissertation, co-authored with Tomas Havranek and Zuzana Irsova, was published as "Tuition Fees and University Enrollment: A Meta-Regression Analysis" in the Oxford Bulletin of Economics and Statistics (Zeynalova *et al.* 2018). A serious debate about the relationship between tuition fees and demand for higher education dates back to Jackson & Weathersby (1975). Although there were occasional large price elasticities in the empirical literature, most evidence suggested that higher education demand is relatively price-inelastic. Researchers have provided various explanations for the observed lack of large elasticities, including the compensating effect of financial aid, the higher earnings of graduates compared to non-graduates, historically small tuition fee increases in real terms, aggressive marketing, students' willingness to pay for quality, the expansion of the student pool with female and minority participants, and the higher income of many university students' families. Our goal was to review the extensive research that has been conducted on this topic, identify patterns in the results, and determine the mean effect that could serve as 'the best estimate' for public policy purposes.

We collected 442 estimates from 43 studies on the relationship between tuition fees and university enrollment. The literature indicates significant publication selection against positive estimates, suggesting that many researchers use the sign of the estimated effect as a specification test to show that education is unlikely a Giffen good. We also found evidence of systematic dependencies between the estimated effects and data, methodological, and publication characteristics. Male students displayed larger tuition elasticities, as did students at private universities. Our results also suggest that the reported relationship is stronger for US students and when panel data are used, while it is weaker when income is controlled for and in the short run. We show that the correlation between tuition and enrollment has remained stable over the last 50 years. The mean effect beyond publication bias is close to zero. When we assign greater weight to the more reliable estimates (those methodologically sound), we obtain a similarly small effect.

Chapter 3 of this dissertation, co-authored with Tomas Havranek, Zuzana Irsova and Lubica Laslopova, was published as "Publication and Attenuation Biases in Measuring Skill Substitution" in the Review of Economics and Statistics (Havranek *et al.* 2023). The paper has two sides to its story: the first involves the empirical application of meta-analysis methods to estimate the elasticity of substitution between skilled and unskilled workers, while the second investigates how publication bias (resulting from the under-reporting of small estimates) and attenuation bias (resulting from measurement error) affect the reported inverse elasticity. The literature suggests a simple average elasticity of 1.8, but individual studies estimating the elasticity show a higher level of disagreement than is often acknowledged in the application of these estimates in calibrations. Elasticities greater than one, which indicate that skilled and unskilled labor are gross substitutes, are more prevalent in the literature and often range from around 4. Elasticities smaller than one, indicating that skilled and unskilled labor are gross complements, are less common but do exist.

Negative estimates of elasticity are indeed inconsistent with the canonical model, and zero or infinite estimates are unintuitive. As few researchers are eager to interpret such estimates, the pattern becomes systematic. Correcting for this bias slashes the mean negative inverse elasticity from -0.6 to the vicinity of zero, and the result holds when we relax the common meta-analysis assumption of conditional independence of estimates and standard errors. While publication bias-corrected estimates stemming from the ordinary least squares method and natural experiments remain close to zero, corrected instrumental variable estimates are around -0.25. The result is consistent with attenuation bias in the literature and implies an elasticity of 4 after correction for both biases. The interplay of the two biases in labor economics evokes Griliches (1977), who finds that in measuring the return to education, attenuation bias almost exactly offsets omitted variable bias (which is often correlated with publication bias via specification searching and p-hacking). In our case, publication bias dominates attenuation bias.

Chapter 4 of this dissertation is an unpublished solo-authored manuscript titled "Expected Returns to Higher Education in Russia after Unified State Exam Reform." The primary goal of the policy reform was to decrease the cost of university admission and increase the efficient distribution of knowledge and skills, especially for those residing in small municipalities and rural areas. I hypothesize that the intended impacts of the reform will be reflected in increased returns to education resulting from the reform, at least to some extent. In order to assess the effects on income and salaries, my study utilizes quasi-experimental methods such as difference-in-differences and propensity score, with the introduction of the exam serving as an exogenous event. The findings suggest that the reform did not have a significant impact on individuals in Moscow who did not obtain higher education diplomas. However, those from other major cities, small towns, and rural areas experienced significant positive effects.

My findings are robust and consistent across various methods of analysis, different control groups, and model specifications. Notably, factors such as family background, regional and federal district location have a significant impact on an individual's salary in Russia. Additionally, an individual's gender, marital status, ethnicity, and parental education also significantly influence their earnings. Women typically earn less than men, married individuals earn more than unmarried ones, and those with a parent possessing a higher education degree earn more on average. However, the impact of parental education on earnings varies across different regions of Russia. In Moscow and smaller towns, the father's education is more crucial, while in St. Petersburg and other major cities, the mother's education level has a more significant impact.

References

Griliches, Z. (1977): "Estimating the returns to schooling: Some econometric problems." *Econometrica* **45(1)**: pp. 1–22.

Hanushek, E. & L. Woessmann (2012): "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation." *Journal of Economic Growth* **17(4)**: pp. 267–321.

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

Jackson, G. A. & G. B. Weathersby (1975): "Individual demand for higher education: A review and analysis of recent empirical studies." *The Journal of Higher Education* **46(C)**: pp. 623–651.

World Bank (2018): "World development report 2018—Learning to realize education's promise." *World Bank report*, The World Bank Group, Washington, DC.

Zeynalova, O., T. Havranek, & Z. Irsova (2018): "Tuition fees and university enrolment: A meta-regression analysis." *Oxford Bulletin of Economics and Statistics* **80(6)**: pp. 1145–1184.

Chapter 2

Tuition Fees and University Enrollment: A Meta-Regression Analysis

Abstract

One of the most frequently examined relationships in education economics is the correlation between tuition fee increases and the demand for higher education. We provide a quantitative synthesis of 443 estimates of this effect reported in 43 studies. While large negative estimates dominate the literature, we show that researchers report positive and insignificant estimates less often than they should. After correcting for this publication bias, we find that the literature is consistent with the mean tuition-enrollment elasticity being close to zero. Nevertheless, we identify substantial heterogeneity among the reported effects: for example, male students and students at private universities display larger elasticities. The results are robust to controlling for model uncertainty using both Bayesian and frequentist methods of model averaging.

Keywords: Enrollment, tuition fees, demand for higher education, meta-analysis, publication bias, model averaging

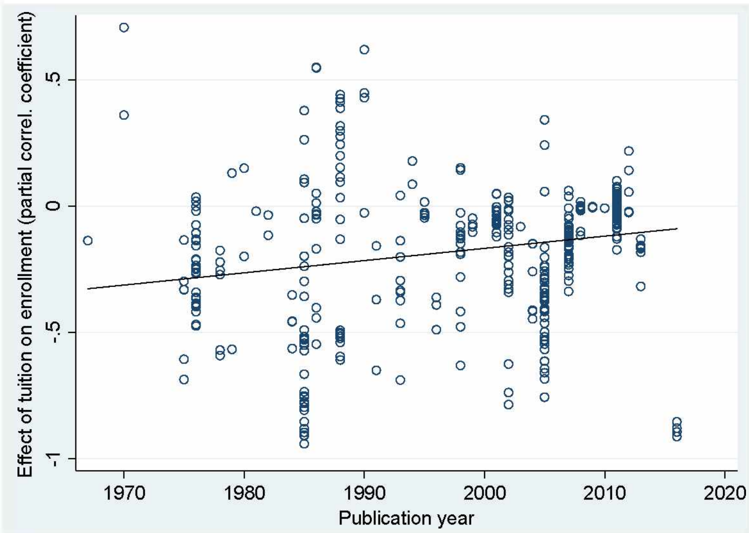
JEL Codes: I23, I28, C52

This paper is a joint work with Tomas Havranek and Zuzana Irsova. The paper is published in the Oxford Bulletin of Economics and Statistics. We thank the editor and four anonymous referees for their useful comments. We are also grateful to Craig Gallet for providing us with the dataset used in his meta-analysis. We acknowledge support from the Czech Science Foundation grant #18-02513S and #16-00027S. This project has also received funding from the European Union's Horizon 2020 Research and Innovation Staff Exchange program under the Marie Sklodowska-Curie grant agreement #681228 and Charles University project PRIMUS/17/HUM/16. Data and code are available in an online appendix at `meta-analysis.cz/education`.

2.1 Introduction

The relationship between the demand for higher education and changes in tuition fees[1] constitutes a key parameter not only for deans but also for policymakers. It is therefore not surprising that dozens of researchers have attempted to estimate this relationship. While the relationship (often, but not always, presented in the form of an elasticity) can be expected to vary somewhat across different groups of students and types of universities, there has been no consensus even on the mean effect, as many literature surveys demonstrate (see, for example, Jackson & Weathersby 1975; Leslie & Brinkman 1987; Heller 1997): the estimates often differ by an order of magnitude, as we also show in Figure 2.1.

Figure 2.1: No clear message in 50 years of research



Notes: The figure depicts a common metric (partial correlation coefficient) of the reported effect of tuition fees on enrollment in higher education institutions. The time trend is not statistically significant.

The academic discussion concerning the correlation between tuition fees and demand for higher education dates back at least to Ostheimer (1953). Even though large price elasticities do occasionally appear in the empirical literature (see, among others, Agarwal & Winkler 1985; Allen & Shen 1999; Buss *et al.* 2004), the majority

[1]For parsimony, in this paper, we usually omit “fee” and use the word “tuition” in its North American sense, “a sum of money charged for teaching by a college or university.”

of the evidence corroborates the notion of a rather price-inelastic demand for higher education across many contexts. Researchers offer numerous explanations for the observed lack of large elasticities: for example, the effect of financial aid compensating tuition changes (Canton & de Jong 2002), increasing earnings of graduates relative to those of non-graduates (Heller 1997), historically small tuition fee increases in real terms and the impact of aggressive marketing (Leslie & Brinkman 1987), larger student willingness to pay for quality (McDuff 2007), expansion of the student pool with female and minority participants, and the fact that many university students come from higher-income families (Canton & de Jong 2002). Even the very first literature review by Jackson & Weathersby (1975) put forward the case for the correlation between tuition and enrollment, while significant and negative, to be rather small in magnitude.

The existing narrative literature surveys, including Jackson & Weathersby (1975), McPherson (1978), Chisholm & Cohen (1982), Leslie & Brinkman (1987), and Heller (1997), place the tuition-enrollment relationship below a 1.5 percentage-point change per $100 tuition increase. The first quantitative review on this topic, Gallet (2007), puts the mean tuition elasticity of demand for higher education at -0.6. However, every single review acknowledges that the mean estimate could be somehow biased and driven by the vast differences in the design of studies, namely, methodological (Quigley & Rubinfeld 1993), country-level (Elliott & Soo 2013), institution-level (Hight 1975), and qualitative differences. Our goal in this paper is to exploit the voluminous work of previous researchers on this topic, assign a pattern to the differences in results, and derive a mean effect that could be used as “the best estimate for public policy purposes” that the literature has sought to identify (Leslie & Brinkman 1987, p. 189).

Achieving our two goals involves collecting the reported estimates of the effect of tuition fees on enrollment and regressing them on the characteristics of students, universities, and other aspects of the data and methods employed in the original studies. Such a meta-analysis approach is complicated by two problems that have yet to be addressed in the literature on tuition and enrollment: publication selection and model uncertainty. Publication selection arises from the common preference of authors, editors, and referees for results that are intuitive and statistically significant. In the context of the tuition-enrollment nexus, one might well treat positive estimates with suspicion as few economists consider education to be Giffen good. Nevertheless, sufficient imprecision in estimation can easily yield a positive estimate, just as it can yield a very large negative estimate. The zero boundary provides a useful rule of thumb for model specification, but the lack of symmetry in the selection rule will typically lead to an exaggeration of the mean reported effect (Doucouliagos & Stanley 2013).

The second problem, model uncertainty, arises frequently in meta-analysis because many factors may influence the reported coefficients. Nevertheless, absent clear guidance that would specify which variables (out of the many dozen potentially useful ones) must be included in and which must be excluded from the model, researchers face a dilemma between model parsimony and potential omitted variable bias. The easiest solution is to employ stepwise regression, but this approach is not appropriate because important variables can be excluded by accident in sequential t-tests (this problem is inevitable, to some extent, also with more sophisticated methods of model selection—every time we need to choose which variables to exclude).[2] In contrast, we employ model averaging techniques that are commonly used in growth regressions: Bayesian model averaging and frequentist model averaging, which are well described and compared by Amini & Parmeter (2012). The essence of model averaging is to estimate (nearly) all models with the possible combinations of explanatory variables and weight them by statistics related to goodness of fit and parsimony.[3]

Our results suggest that the mean reported relation between tuition and enrollment is significantly downward biased because of publication selection (in other words, positive and insignificant estimates of the relationship are discriminated against).

[2] Campos *et al.* (2005) provide a useful review of general-to-specific modeling.

[3] Model averaging allows us to take into account the model uncertainty associated with our meta-analysis model. Nevertheless, this approach does not address the model uncertainty in estimating the tuition-enrollment nexus in primary studies: this second source of model uncertainty is the reason for conducting a meta-analysis in the first place (Stanley & Jarrell 1989; Stanley & Doucouliagos 2012). A technical treatment of these two sources of model uncertainty with relation to Bayesian model averaging is available in Appendix 2B of Havranek *et al.* (2017).

After correcting for publication selection, we find no evidence of a tuition-enrollment nexus on average. This result holds when we construct a synthetic study with ideal parameters (such as a large dataset, control for endogeneity, etc.) and compute the implied "best-practice estimate": this estimate is also close to zero. Nevertheless, we find evidence of substantial and systematic heterogeneity in the reported estimates. Most prominently, our results suggest that male students and students at private universities display substantial tuition elasticities.

The paper is organized as follows. Section 2.2 describes our approach to data collection and the basic properties of the dataset. Section 2.3 tests for the presence of publication selection bias. Section 2.4 explores the data, method, and publication heterogeneity in the estimated effects of tuition fees on enrollment and constructs a best practice estimate of the relationship. Section 2.5 provides extensions and robustness checks. Section 2.6 concludes the paper. An online appendix, available at meta-analysis.cz/education, provides the data and code that allow other researchers to replicate our results.

2.2 The Dataset

Researchers often, but not always, estimate the tuition-enrollment relationship in the form of the price elasticity of demand for higher education:

$$\ln Enrollment_{it} = \alpha + PED \cdot \ln Tuition_{it} + YED \cdot \ln Income_{it} + Controls_{ijt} + \epsilon_{it}, \tag{2.1}$$

where the demand for education $Enrollment_{it}$ typically denotes the total number of students enrolled in higher education institution i in time period t, *Tuition* denotes the tuition payment for higher education, *Income* denotes the family income of a student, and its respective coefficient *YED* denotes the income elasticity of demand. ϵ is the error term. The vector $Controls_{ijt}$ represents a set of explanatory variables j, such as proxies for the quality of education (university ranking, percentage of full professors employed, student/faculty ratio, average score on assessment tests), funding opportunities (grants, external financial support, the cost of loans), or labor market conditions (the level of unemployment or the wage gap between university-educated and high school-educated workers).

From the empirical literature reporting the correlation between tuition fees and the demand for higher education, we collect the coefficient *PED*. In (2.1), *PED* denotes the elasticity and captures the percent change in demand for higher education if tuition increases by one percent. The relationship between enrollment and tuition is, however, not always estimated in the literature in the form of an elasticity; sometimes other versions of (2.1) than log-log are used: the relationship can be linear or represented by the student price response coefficient (Jackson & Weathersby 1975). Moreover, the definitions of the tuition and enrollment variables vary: while tuition can represent net financial aid or include other fees, enrollment can represent the total headcount of the enrolled, the number of applications, the percentage of enrolled students, or enrollment probability. Even the uncertainty measure surrounding the point estimates reported in the literature cannot always be converted into a standard error.

To be able to focus solely on elasticities and simultaneously make the sample fully comparable, we would need to eliminate a substantial part of the data (just as Gallet 2007, did; moreover, our study faces an additional sample reduction since not all studies report an uncertainty measure, which we need to account for in estimating publication bias). Maximizing the number of observations and minimizing the mistakes made through conventional conversion calls for a different type of common metric. McPherson (1978, p. 180) supports the case of an ordinal measure: "*There is probably not a single number in the whole enrollment demand literature that should be taken seriously by itself. But a careful review of the literature will show that there are some important qualitative findings and order-of-magnitude estimates on which there is consensus, and which do deserve to be taken seriously.*" Therefore, we use all estimates of the tuition-enrollment nexus, including linear and semi-log specifications. We follow Doucouliagos (1995), Djankov & Murrell (2002), Doucouliagos & Laroche (2003), Babecky & Havranek (2014), Valickova *et al.* (2015), and Havranek *et al.* (2016), among others, and convert the collected estimates into partial correlation coefficients, which transform t-values to a measure that is not related to the size of the dataset. Now, the PED coefficient is standardized to

$$PCC(PED)_{ij} = \frac{T(PED)_{ij}}{\sqrt{T(PED)_{ij}^2 + DF(PED)_{ij}}}, \tag{2.2}$$

where $PCC(PED)_{ij}$ represents the estimated partial correlation coefficient of the i-th estimate of the tuition elasticity PED, with $T(PED)_{ij}$ representing the corresponding t-statistics and $DF(PED)_{ij}$ representing the corresponding number of degrees of freedom reported in the j-th study. We take advantage of the previously published surveys on this topic, especially Leslie & Brinkman (1987), Heller (1997), and Gallet (2007), and extend the data sample by searching the Google Scholar database. The search query is available online at meta-analysis.cz/education. We added the last study on September 23, 2016.

The sample of studies we collect is subjected to three major selection criteria. First, the study must investigate the relationship between tuition and enrollment with enrollment as the dependent variable. This criterion eliminates multiple studies, including Mattila (1982), Galper & Dunn (1969), and Christofides *et al.* (2001), which estimate only income effects on enrollment. Second, the explanatory variable *Tuition* cannot be a dummy variable, which excludes studies such as Bruckmeier & Wigger (2014) and Dwenger *et al.* (2012) (Hübner 2012, for example, uses a dummy variable indicating residence in a fee state to investigate the effects of tuition on enrollment probabilities). Third, the study must report a measure of uncertainty around the estimate (Corman & Davidson 1984, for example, report neither t-statistics nor standard errors). The final sample of studies used in our meta-analysis is listed in Table 2.1.

Previous literature surveys argue for a relatively modest magnitude of the relationship between tuition and enrollment (generally in terms of the mean student price response coefficient): Jackson & Weathersby (1975), a survey of 7 studies published between 1967 and 1973, places the enrollment change in the range of $(-0.05, -1.46)$ percentage points per \$100 tuition increase in 1974 dollars; McPherson (1978) up-

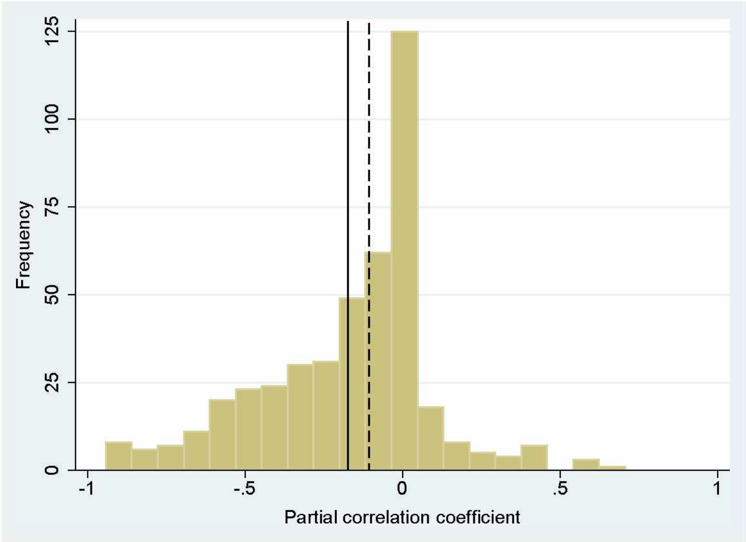
Table 2.1: Studies used in the meta-analysis

Agarwal & Winkler (1985)	Doyle & Cicarelli (1980)	Murphy & Trandel (1994)
Alexander & Frey (1984)	Elliott & Soo (2013)	Noorbakhsh & Culp (2002)
Allen & Shen (1999)	Grubb (1988)	Ordovensky (1995)
Berger & Kostal (2002)	Hemelt & Marcotte (2011)	Parker & Summers (1993)
Bezmen & Depken (1998)	Hight (1975)	Paulsen & Pogue (1988)
Bruckmeier *et al.* (2013)	Hoenack & Pierro (1990)	Quigley & Rubinfeld (1993)
Buss *et al.* (2004)	Hoenack & Weiler (1975)	Quinn & Price (1998)
Campbell & Siegel (1967)	Hsing & Chang (1996)	Savoca (1990)
Canton & de Jong (2002)	Huijsman *et al.* (1986)	Shin & Milton (2008)
Chen (2016)	Kane (2007)	Suloc (1982)
Cheslock (2001)	King (1993)	Tannen (1978)
Chressanthis (1986)	Knudsen & Servelle (1978)	Toutkoushian & Hollis (1998)
Coelli (2009)	Koshal *et al.* (1976)	Tuckman (1970)
Craft *et al.* (2012)	McPherson & Schapiro (1991)	
Dearden *et al.* (2011)	Mueller & Rockerbie (2005)	

dates the range to $(-0.05, -1.53)$. Leslie & Brinkman (1987), a survey of 25 studies published between 1967 and 1982, places the mean student price response coefficient at -0.7 per \$100 in 1982 dollars; and Heller (1997), a survey of 8 studies published between 1990 and 1996, reports a range of $(-0.5, -1.0)$. The first literature survey to examine quantitatively the heterogeneity in the estimates appears much later: Gallet (2007), a meta-analysis of 295 observations from 53 studies published between 1953 and 2004, reports a mean tuition elasticity of demand for higher education of -0.6.

Our final dataset covers 43 studies comprising 442 estimates of the relationship between enrollment in a higher education institution and tuition recalculated to partial correlation coefficients. The oldest study was published in 1967, and the newest was published in 2016, representing half of a century of research in the area. The (left-skewed) distribution of the reported coefficients is shown in Figure 2.2; the coefficients range from $-.941$ to $.707$ and are characterized by a mean of -0.171 and a median of -0.103. Approximately 25% of the estimates are larger than 0.33 in the absolute value, which, according to Doucouliagos (2011), can be classified as a "large" partial correlation coefficient, while the mean coefficient is classified as a borderline "medium" effect. Nevertheless, using Cohen's guidelines for correlations in social sciences (Cohen 1988), the mean effect of -0.171 would be classified as a "small" effect. Although the histogram only has one peak, Figure 2.5 and Figure 2.6 (presented in the Appendix) suggest substantial study- and country-level heterogeneity.

Figure 2.2: Histogram of the partial correlation coefficients



Notes: The figure depicts a histogram of the partial correlation coefficients of the enrollment-tuition nexus estimates reported by individual studies. The dashed vertical line denotes the sample median, and the solid vertical line denotes the sample mean.

Consequently, we collect 17 explanatory variables that describe the data and model characteristics and investigate the possible reasons for heterogeneity below in Section 2.4.

Table 2.2 provides us with some preliminary information on the heterogeneity in the estimates. Estimating the mean partial correlation via restricted maximum likelihood random effects meta-analysis using the Hartung-Knapp modification (presented in the last row of the table) does not affect our previous discussion regarding the average size of the effect in question. Next, we summarize the simple mean values for each category and mean values weighted by the inverse number of estimates reported per study (to assign each study the same weight) according to different data, methodological, and publication characteristics. Larger studies with many estimates largely drive the simple mean of the partial correlation coefficients, especially in samples that consider private universities, female students, and countries outside the US (a large portion of the estimates in the literature are estimated data from US universities, but some studies focus on other countries, especially in Europe). Thus, it seems reasonable to focus on the weighted statistics in the following discussion.

Table 2.2: Partial correlation coefficients for different subsets of data

		Unweighted			Weighted		
	Obs.	Mean	95% conf. int.		Mean	95% conf. int.	
Temporal dynamics							
Short-run effect	209	-0.106	-0.134	-0.078	-0.135	-0.169	-0.101
Long-run effect	233	-0.229	-0.269	-0.189	-0.233	-0.277	-0.189
Estimation technique							
Control for endogeneity	31	-0.034	-0.144	0.076	-0.043	-0.135	0.050
No control for endogeneity	411	-0.181	-0.207	-0.155	-0.219	-0.249	-0.189
Data characteristics							
Private universities	115	-0.086	-0.127	-0.044	-0.236	-0.284	-0.188
Public universities	160	-0.198	-0.243	-0.153	-0.154	-0.207	-0.101
Male candidates	49	-0.330	-0.389	-0.270	-0.329	-0.395	-0.263
Female candidates	46	-0.252	-0.318	-0.186	-0.165	-0.220	-0.110
Spatial variation							
USA	355	-0.150	-0.179	-0.121	-0.196	-0.229	-0.162
Other countries	87	-0.256	-0.305	-0.207	-0.136	-0.171	-0.102
Publication status							
Published study	262	-0.249	-0.288	-0.211	-0.209	-0.247	-0.170
Unpublished study	180	-0.056	-0.075	-0.038	-0.076	-0.106	-0.047
Publication year							
Until 1980	48	-0.227	-0.299	-0.155	-0.206	-0.308	-0.103
1981–1990	80	-0.246	-0.342	-0.151	-0.129	-0.215	-0.044
1991–2000	44	-0.191	-0.256	-0.127	-0.191	-0.259	-0.123
2001–2010	144	-0.218	-0.254	-0.183	-0.208	-0.245	-0.170
Since 2011	126	-0.040	-0.069	-0.011	-0.206	-0.263	-0.148
All estimates	442	-0.171	-0.196	-0.145	-0.186	-0.214	-0.158
Random effects MA	442	-0.156	-0.180	-0.131	-0.167	-0.192	-0.142

Notes: The table reports mean values of the partial correlation coefficients for different subsets of data. The exact definitions of the variables are available in Table 2.4. Weighted = estimates that are weighted by the inverse of the number of estimates per study. MA = meta-analysis.

We observe differences between the short- and long-term effects, which appear to be in line with intuition: a larger negative long-term coefficient would suggest that in the long run, students have more time to search for other competing providers of education. A substantial difference also appears when researchers do not account for the presence of endogeneity in the demand equation: controlling for endogeneity diminishes the partial correlation coefficient by 0.15; the effect itself is on the boundary between a small and medium effect, according to Doucouliagos's guidelines.

The evidence on one of the most widely studied topics in the literature, the difference in the elasticity between public and private institutions, changes when weighting is applied: Hopkins (1974), for example, finds that students in private institutions have a higher elasticity than those in public universities, which is consistent with the weighted average from Table 2.2. The simple average is to some extent skewed by the considerable number of positive estimates in larger studies (Grubb 1988; Hemelt & Marcotte 2011), which would correspond to a situation in which private universities use tuition as a signal of the quality of the university. Male candidates seem to display a larger elasticity to changes in tuition fees than female candidates, and the difference increases when weighting is applied (the result is well in line with Huijsman *et al.* 1986, but contradicts Bruckmeier *et al.*, 2013, who do not find any differences). Spatial differences do not seem to be extensive; however, we observe that published studies report larger estimates of the effect of tuition on demand for higher education. Table 2.2 also shows that the estimates do not vary much in time (the apparent drop in the estimates over the last decade disappears when estimates are weighted). The differences in results between published and unpublished studies might indicate the presence of publication bias, although not necessarily.

2.3 Publication Bias

Publication selection bias is especially likely to occur when there is a strong preference in the literature for a certain type of result. Both editors and researchers often yearn for significant estimates of a magnitude consistent with the commonly accepted theory. The law of demand, which implies a negative relationship between the price and demanded quantity of a good, is taken to be one of the most intuitive economic relationships; education is unlikely to be perceived as a Giffen good (Doyle & Cicarelli 1980).[4] Therefore, researchers may treat positive estimates of the tuition-enrollment nexus with suspicion and sometimes do explicitly refer to the conventional expectation of the desired sign. (Canton & de Jong 2002, p. 657), for example, comment on their results as follows: *"We find that the short-run coefficients*

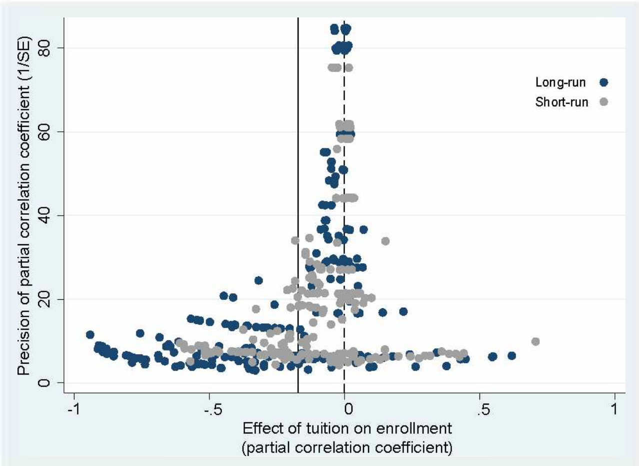
[4]On the other hand, it is important to mention the potential snob effect related to the price of education and anticipated in the discussion of Table 2.2. While the association between prices and demand is unlikely to be positive on average, it might easily be positive for some individual schools, students, or parents. This observation suggests that in the case of higher education, positive price elasticities may be somewhat more acceptable than, for example, those in the literature on gasoline demand (Havranek *et al.* 2012).

all have the 'right' sign, except for the positive but insignificant coefficient on tuition fees ..."

Indeed, the unintuitive sign of an estimate might indicate identification problems; the probability of obtaining the 'wrong' sign increases with small samples, noisy data, or misspecification of the demand function (Stanley 2005). We should, however, obtain the unintuitive sign of an estimate from time to time just by chance. Systematic under-reporting of estimates with the 'wrong' sign drives the global mean in the opposite direction. This distortion of reported results is a frequently reported phenomenon in economic research (for example, among other studies, Doucouliagos & Stanley 2013; Ioannidis *et al.* 2017; Havranek & Irsova 2011; 2012; Rusnak *et al.* 2013; Havranek & Kokes 2015; Havranek *et al.* 2015b). Studies addressing the law of demand are frequently affected by publication selection, but other areas also suffer from bias, with the economics of education being no exception: Fleury & Gilles (2015) report publication bias in the literature on the inter-generational transmission of education, Ashenfelter *et al.* (1999) find bias in the estimates of the rate of return to education, and Benos & Zotou (2014) report bias toward a positive impact of education on growth. Primary studies could, of course, incorporate theoretical expectations about the elasticity formally as priors within a Bayesian estimation framework, but this approach is unfortunately not used in the literature on the tuition-enrollment nexus.

The so-called funnel plot commonly serves as a visual test for publication bias (see, for example, Stanley & Doucouliagos 2010, and the studies cited therein). It is a scatter plot with the effect's magnitude on the horizontal axis and its precision (the inverse of the standard error) on the vertical axis (Stanley 2005). In the absence of publication bias, the graph resembles an inverted funnel, with the most precise estimates close to the underlying effect; with decreasing precision, the estimated coefficients become more dispersed and diverge from the underlying effect. Moreover, if the coefficients truly estimate the underlying effect with some random error, the inverted funnel should be symmetrical. The asymmetry in Figure 2.3 is consistent with the presence of publication bias related to the sign of the effect; if the bias is

Figure 2.3: The funnel plot suggests publication selection bias



Notes: The dashed vertical line indicates a zero partial correlation coefficient of the elasticity of demand for higher education; the solid vertical line indicates the mean partial correlation coefficient. When there is no publication selection bias, the estimates should be symmetrically distributed around the mean effect.

related to statistical significance, the funnel becomes hollow and wide. The literature exhibits a very similar pattern of bias for the short- and long-term elasticity estimates; thus, in the calculations that follow, we do not further divide the sample based on these two characteristics, but we control for the differences in the next section.

Following Stanley (2005) and Stanley (2008), we examine the correlation between the partial correlation coefficients PCCs and their standard errors in a more formal, quantitative way:

$$PCC_{ij} = PCC_0 + \beta \cdot SE(PCC_{ij}) + \mu_{ij}, \tag{2.3}$$

where PCC_{ij} denotes i-th effect with the standard error $SE(PCC_{ij})$ estimated in the j-th study and μ_{ij} is the error term. The intercept of the equation, PCC_0, is the 'true' underlying effect absent publication bias; the coefficient of the standard error, β, represents publication bias. In the case of zero publication bias ($\beta = 0$), the estimated effects should represent an underlying effect that includes random error. Otherwise ($\beta \neq 0$), we should observe correlation between the PCCs and their standard error, either because researchers discard positive estimates of PCCs ($\beta < 0$) or because researchers compensate large standard errors with large estimates

Table 2.3: Funnel asymmetry tests detect publication selection bias

Panel A: Unweighted sample	OLS	IV	Proxy	Median
SE (publication bias)	-1.142***	-1.915***	-1.523***	-1.318**
	(0.36)	(0.43)	(0.27)	(0.52)
Constant (effect absent bias)	-0.059	0.016	-0.004	-0.035
	(0.06)	(0.04)	(0.04)	(0.07)
Observations	442	442	442	442
Panel B: Weighted sample	Precision		Study	
	WLS	IV	OLS	IV
SE (publication bias)	-1.757***	-2.305***	-1.023***	-1.753***
	(0.36)	(0.49)	(0.29)	(0.52)
Constant (effect absent bias)	0.001	0.026	-0.069**	0.015
	(0.02)	(0.02)	(0.03)	(0.05)

Notes: The table reports the results of the regression $PCC_{ij} = PCC_0 + \beta \cdot SE(PCC_{ij}) + \mu_{ij}$, where PCC_{ij} denotes i-th tuition elasticity of demand for higher education estimated in the j-th study and $SE(PCC_{ij})$ denotes its standard error. Panel A reports results for the whole sample of estimates, and Panel B reports the results for the whole sample of estimates weighted by precision or study. OLS = ordinary least squares. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. Proxy = the inverse of the square root of the number of observations is used as a proxy for the standard error. Median = only median estimates of the tuition elasticities reported in the studies are included. Study = model is weighted by the inverse of the number of estimates per study. Precision = model is weighted by the inverse of the standard error of an estimate. WLS = weighted least squares. Standard errors in parentheses are robust and clustered at the study and country level (two-way clustering follows Cameron *et al.* 2011). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

of PCCs. In other words, the properties of the standard techniques used to estimate the tuition-enrollment nexus yield a t-distribution of the ratio of point estimates to their standard errors, which means that the estimates and standard errors should be statistically independent quantities.

Table 4.6 reports the results of (2.3). In Panel A, we present four different specifications applied to the unweighted sample: simple OLS, an instrumental variable specification in which the instrument for the standard error is the inverse of the square root of the number of observations (as in, for example, Stanley 2005; Havranek *et al.* 2018b); OLS, in which the standard error is replaced by the aforementioned instrument (as in Havranek 2015); and study-level between-effect estimation.[5] In Panel B, we weight all estimates by their precision, which assigns greater importance to more precise results and directly corrects for heteroskedasticity (Stanley & Jarrell 1989), and furthermore, we weight all estimates by the inverse of the number of observations

[5]It is worth noting that while we also intended to use study-level fixed effects (a common robustness check accounting for unobserved study-level characteristics, see the online appendix of Havranek & Irsova 2017), the high unbalancedness of our panel dataset and the fact that a number of studies report only one observation make this specification infeasible.

per study, which treats small and large studies equally. In accordance with the mean statistics from Table 2.2, the mean effect marginally increases (the mean relation between enrollment and tuition changes becomes less sensitive) and even becomes significant but is still close to zero.

Two important findings can be distilled from Table 4.6. First, publication bias is indeed present in our sample; according to the classification of Doucouliagos & Stanley (2013), the magnitude of selectivity ranges from substantial ($-2 > \beta > -1$) to severe ($\beta < -2$). Second, we cannot reject the hypothesis that the underlying tuition-enrollment effect corrected for publication bias is zero. The estimated coefficient β suggests that the true effect is very small or indeed zero. Nevertheless, Table 4.6 does not tell us whether data and method choices are correlated with the magnitude of publication bias or the underlying effect. We address these issues in the next section.

2.4 Heterogeneity

2.4.1 Variables and Estimation

Thirty years ago, Leslie & Brinkman (1987) concluded their review of the tuition-enrollment literature with disappointment regarding study heterogeneity: *"Weinschrott (1977) was correct when he warned about the difficulties in achieving consistency among such disparate studies."* Data heterogeneity in our own sample is obvious from Figure 2.5 and Figure 2.6, presented in the Appendix, and the substantial standard deviations of the mean statistics we report in Table 2.2. Therefore, we code 17 characteristics of study design as explanatory variables that capture additional variation in the data. The explanatory variables are listed in Table 2.4 and divided into four groups: variables capturing methodological differences, differences in the design of the demand function, differences in the dataset, and publication characteristics. Table 2.4 also includes the definition of each variable, its simple mean, standard deviation, and the mean weighted by the inverse of the number of observations extracted from a study.

Table 2.4: Description and summary statistics of regression variables

Variable	Description	Mean	SD	WM
Partial correlation coef.	Partial correlation coefficient derived from the estimate of the tuition-enrollment relationship.	-0.171	0.271	-0.186
Standard error	The estimated standard error of the tuition-enrollment estimate.	0.097	0.070	0.115
Estimation characteristics				
Short-run effect	= 1 if the estimated tuition-enrollment effect is short-term (in differences) instead of long-term (in levels).	0.473	0.500	0.480
OLS	= 1 if OLS is used for the estimation of the tuition-enrollment relationship.	0.446	0.498	0.687
Control for endogeneity	= 1 if the study controls for price endogeneity.	0.070	0.256	0.187
Design of the demand function				
Linear function	= 1 if the functional form of the demand equation is linear.	0.296	0.457	0.301
Double-log function	= 1 if the functional form of the demand equation is log-log.	0.507	0.501	0.501
Unemployment control	= 1 if the demand equation controls for the unemployment level.	0.495	0.501	0.350
Income control	= 1 if the demand equation controls for income differences.	0.643	0.480	0.653
Data specifications				
Cross-sectional data	= 1 if cross-sectional data are used for estimation instead of time-series or panel data.	0.204	0.403	0.303
Panel data	= 1 if panel data are used for estimation instead of cross-sectional or time-series data.	0.557	0.497	0.347
Male candidates	= 1 if the study estimates the tuition-enrollment relationship for male applicants only.	0.111	0.314	0.075
Female candidates	= 1 if the study estimates the tuition-enrollment relationship for female applicants only.	0.104	0.306	0.051
Private universities	= 1 if the study estimates the tuition-enrollment relationship for private universities only.	0.260	0.439	0.233
Public universities	= 1 if the study estimates the tuition-enrollment relationship for public universities only.	0.362	0.481	0.454
USA	= 1 if the tuition-enrollment relationship is estimated for the United States only.	0.803	0.398	0.839
Publication characteristics				
Publication year	Logarithm of the publication year of the study.	7.601	0.006	7.598
Citations	Logarithm of the number of citations the study received in Google Scholar.	3.529	1.079	3.227
Published study	= 1 if the study is published in a peer-reviewed journal.	0.593	0.492	0.828

Notes: SD = standard deviation, SE = standard error, WM = mean weighted by the inverse of the number of estimates reported per study.

Estimation characteristics: The exact distinction between short- and long-run effects is disputable in most economic studies (see, for example Espey 1998). If the author does not clearly designate her estimate, we follow the basic intuition and classify the growth estimates as short-term and the level estimates as long-term. Static models, however, introduce ambiguity. If the dataset covers only a short period of time, the estimate might not reflect the full long-term elasticity; thus, we label such estimates as *short-run effects*. Hoenack (1971) notes the importance of temporal dynamics: lowering costs in the long run encourages students to apply for higher education; in the short run, however, the change can only influence the current applicants. The long-run effects are therefore likely to be larger. We do not divide the sample between short- and long-run elasticities, which conforms to our previous discussion and the practice applied by the previous meta-analyses on this topic (Gallet 2007).

Researchers use various techniques to estimate the tuition-enrollment relationship. Fixed effects in particular dominate the panel-data literature. More than one-third of the estimates are a product of simple *OLS*, and surprisingly few studies control for endogeneity: as Coelli (2009) emphasizes, an increase in tuition fees could be a response to an increase in the demand for higher education. Therefore, the estimated coefficient may also include a positive price response to the supply of student vacancies and thus underestimate the effect of tuition on the demand for higher education (Savoca 1990). For this reason, we could expect the estimates that do not account for the endogeneity of tuition fees, such as those derived using OLS, to indicate a correlation with enrollment different than those derived using, say, instrumental variables (as in Neill 2009, for example).[6] To address this endogeneity bias, we include a dummy variable indicating methods that do *control for endogeneity*.

Design of the demand function: The relationship between tuition fees and the demand for higher education can be captured in multiple ways. We present in (2.1)

[6]Note that some studies, such as Coelli (2009), use OLS while simultaneously attempting to minimize endogeneity bias using other than methodological treatments, mostly by including a detailed set of individual youth and parental characteristics.

the *double-log functional form* of the demand function, which produces the elasticity measure and accounts for half of the estimates in our sample (Allen & Shen 1999; Noorbakhsh & Culp 2002; Buss *et al.* 2004). Some authors, including McPherson & Schapiro (1991) and Bruckmeier *et al.* (2013), capture the simple linear relationship between the variables using a *linear demand function*. Semi-elasticities are also sometimes estimated and can be captured by the semi-log functional form (Shin & Milton 2008); several authors use non-linear Box-Cox transformations (such as Hsing & Chang 1996, who test whether the estimated elasticity is indeed constant). We suspect that despite the transformation of all the estimates into partial correlation coefficients, some systematic deviations in the estimates might remain based on the form of the demand function.

Researchers also specify demand equations to reflect various social and economic conditions of the applicants. We account for whether researchers control for the two most important of these conditions: the *income* level and *unemployment* rate. Lower-income students should be more responsive to changes in tuition than higher-income students (McPherson & Schapiro 1991); we expect systematic differences between results that do and do not account for income differences. The effect of controlling for the unemployment level is not as straightforward. Some authors (such as Berger & Kostal 2002) hypothesize that the unemployment rate might be positively associated with enrollment, as attending a higher education institution can represent a substitute for being employed. An unfavorable employment rate, by contrast, reduces the possibilities of financing higher education. Labor market conditions can also be captured by other variables, for example, real wages, as the opportunity costs of attending university (Mueller & Rockerbie 2005) or a wage gap (Bruckmeier *et al.* 2013) reflecting the differences in earnings between those who did and did not participate in higher education.

Data specifications: Leslie & Brinkman (1987) note that while cross-sectional studies reflect the impact of explicit prices charged in the sample, panel studies reflect that each educational institution implicitly accounts for the price changes of other institutions. Different projection mechanisms could introduce heterogeneity in the estimates. Thus, we include a dummy variable for studies that rely on *cross-sectional* variation and for studies that rely on *panel data* (the reference category being time-series data). Since approximately 80% of our data are estimates for the *USA*, we plan to examine whether geography induces systematic differences in the estimated partial correlation coefficients. Elliott & Soo (2013) conduct a study of 26 different countries including the US: the global demand for higher education seems to be more price sensitive than US demand, although this conclusion is not completely robust.

The issue of *male* and *female* participants and their elasticities with respect to price changes has also been discussed in previous studies. Savoca (1990) claims that females could face lower earnings upon graduation; therefore, they may see higher education as a worse investment and be less likely to apply. Bruckmeier *et al.* (2013) shows that gender matters when technical universities are considered, while Mueller & Rockerbie (2005) find that male Canadian students are more price sensitive than their female counterparts. McPherson & Schapiro (1991), however, argue that the gender effect is in general constant across income groups, and Gallet (2007) does not find significant gender-related differences in reported estimates.

The differences between *public* and *private* educational institutions are also frequently discussed, and researchers agree that these institutions face considerably different demand unless student aid is provided. The results of Funk (1972) suggest the student price response to be consistently lower for private universities. Hight (1975) supports these conclusions and argues that the demand for community or public colleges tends to be more elastic than the demand for private colleges. In a similar vein, Leslie & Brinkman (1987) note that the average student at a private university has a higher family income base; furthermore, a lower-income student, who is also more likely to enroll in a public university, typically demonstrates higher tuition elasticities. However, Bezmen & Depken (1998) find those who apply to private universities to be more price sensitive.

Publication characteristics: While we do our best to control for the relevant data and method features, it is unfeasible to codify every single difference among all estimates. There might be unobserved aspects of data and methodology (or, more generally, quality) that drive the results. For this reason, a number of modern meta-analyses (such as Havranek *et al.* 2015a) employ a variable representing the *publication year* of the study: new studies are more likely to present methodological innovations that we might have missed in our previous discussion. Moreover, the equlibrium elasticity might have changed over time. It is plausible to argue that earlier in the sample, higher education is more or less a luxury good. More recently, however, with increasing higher education enrollment, higher education might have become more of a necessity. Furthermore, we exploit the *number of citations* in Google Scholar to reflect how heavily the study is used as a reference in the literature and information on *publication status* since the peer-review process can be thought of as an indication of study quality.

The purpose of this section is to investigate which of the method choices systematically influence the estimated partial correlation coefficients and whether the estimated coefficient of publication bias from Section 2.3 survives the addition of these variables. Ideally, we would like to regress the partial correlation coefficient on all 17 characteristics listed above, plus the standard error. Since we have a relatively large number of explanatory variables, however, it is highly probable that some of the variables will prove redundant. The traditional use of model selection methods (such as eliminating insignificant variables one by one or choosing the final model specification in advance) often leads to overly optimistic confidence intervals. In this paper, we opt for model averaging techniques, which can address the model uncertainty inherent in meta-analysis.

Bayesian model averaging (BMA) is our preferred choice of estimation technique to analyze heterogeneity. BMA processes hundreds of thousands of regressions consisting of different subsets of the 18 explanatory variables. With such a large model space (2^{18} models to be estimated), we decide to follow some of the previous meta-analyses (such as Havranek & Rusnak 2013; Irsova & Havranek 2013; Havranek *et al.*

2018a, who also use the `bms` R package by Feldkircher & Zeugner, 2009) and apply the Markov chain Monte Carlo algorithm, which considers only the most important models. Bayesian averaging computes weighted averages of the estimated coefficients (posterior means) across all the models using posterior model probabilities (analogous to information criteria in frequentist econometrics) as weights. Thus, all the coefficients have an approximately symmetrical distribution with a posterior standard deviation (analogous to the standard error). Each coefficient is also assigned a posterior inclusion probability (analogous to statistical significance), which is a sum of posterior model probabilities for the models in which the variable is included. Further details on BMA can be found, for example, in Eicher *et al.* (2011).

When applying BMA, researchers have to make several choices. The first choice, as we already mentioned, is whether to compute all models or to use the Markov chain Monte Carlo approximation. Generally, with more than 15 variables, it becomes infeasible to compute all models using a standard personal computer, so researchers typically approximate the whole model space by using the model composition Markov chain Monte Carlo algorithm (Madigan & York 1995), which only traverses the most important part of the model space: that is, the models with high posterior model probabilities. The second choice is the weight of the prior on individual coefficients, the g-prior. The priors are almost always set at zero, which is considered to be the safest choice, unless we have a very strong reason to believe that the coefficients should have a particular magnitude (this is not the case in our study). The most commonly used weight gives the prior the same importance as one individual observation: that is, very little. This is called the unit information prior (UIP), and we apply it following Eicher *et al.* (2011). The third choice concerns the prior on model probability. Again, the most commonly used prior simply reflects that we have little knowledge ex ante, and so each model has the same prior weight. Eicher *et al.* (2011) show that the combination of this uniform model prior and the unit information g-prior performs well in predictive exercises. More technical details about BMA in meta-analysis can be found in Appendix B of Havranek *et al.* (2017).

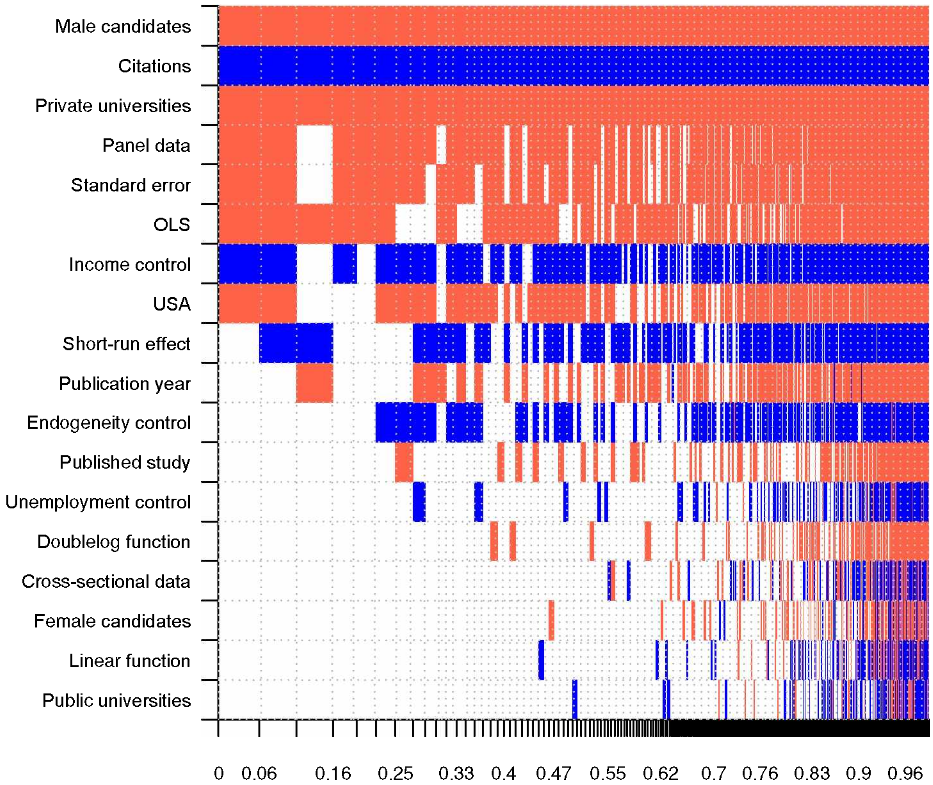
Although BMA is the most frequently used tool to address model uncertainty, recently proposed statistical routines for frequentist model averaging (FMA) make the latter a competitive alternative. Frequentist averaging, unlike the Bayesian version, does not require the use of explicit prior information. We follow Havranek *et al.* (2017), the first study to apply FMA in the meta-analysis framework, who use the approach of Amini & Parmeter (2012), which is based on on the works of Hansen (2007) and Magnus *et al.* (2010). As in the case of BMA, we attempt to restrict our model space from the original 2^{18} models and use Mallow's model averaging estimator (Hansen 2007) with an orthogonalization of the covariate space according to Amini & Parmeter (2012) to narrow the number of estimated models. Mallow's criterion helps to select asymptotically optimal weights for model averaging. Further details on this method can be found in Amini & Parmeter (2012).

2.4.2 Results

The results of the BMA estimation are visualized in Figure 2.4. The rows in the figure represent explanatory variables and are sorted according to the posterior inclusion probability from top to bottom in descending order. The columns represent models and are sorted according to the model inclusion probability from left to right in descending order. Each cell in the figure thus represents a specific variable in a specific model; a blue cell (darker in grayscale) indicates that the estimated coefficient of a variable is positive, a red cell (lighter in grayscale) indicates that the estimated coefficient of a variable is negative, and a blank cell indicates that the variable is not included in the model. Figure 2.4 also shows that nearly half of the variables are included in the best model and that their signs are robustly consistent across different models.

A numerical representation of the BMA results can be found in Table 2.5 (our preferred specification is BMA estimated with the uniform model prior and unit information prior following Eicher *et al.* 2011). Additionally, we provide two alternative specifications: first, a frequentist check estimated by simple OLS with robust standard errors clustered at the study and country level in which we include only variables from BMA with posterior inclusion probability higher than 0.5. Second, we

Figure 2.4: Model inclusion in Bayesian model averaging



Notes: The figure depicts the results of BMA. On the vertical axis, the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of the cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 2.5. All variables are described in Table 2.4. The results are based on the specification weighted by the number of estimates per study.

provide a robustness check based on FMA, which includes all explanatory variables. All estimations are weighted using the inverse of the number of estimates reported per study. In the next section, we also provide the robustness checks of BMA with different priors (following Fernandez *et al.* 2001; Ley & Steel 2009) and different weighting (by precision). Complete diagnostics of the BMA exercises can be found in Section 2.B.

In interpreting the posterior inclusion probability, we follow Jeffreys (1961). The author categorizes values between 0.5 and 0.75 as weak, values between 0.75 and 0.95 as positive, values between 0.95 and 0.99 as strong, and values above 0.99 as decisive evidence for an effect. Table 2.5 thus testifies to decisive evidence of an effect in the cases of *Male candidates*, *Private universities*, and *Citations*; to positive evidence of an effect in the case of *Panel data*; and to weak evidence of an effect in the cases of the *Short-run effect*, *OLS*, *Income control*, and *USA* variables. While our robustness checks seem to support the conclusions from BMA, the evidence for an effect of the variables *OLS* and *Control for endogeneity* changes when FMA is employed.

Publication bias and estimation characteristics: Although diminished to almost half of its original value (Table 4.6), the evidence for publication bias represented by the coefficient on the *Standard error* variable survives the inclusion of controls for data and method heterogeneity. The result supports our original conclusion that publication bias indeed plagues the literature estimating the relationship between tuition fees and the demand for higher education. The evidence on the *short-run effect* is in line with expectation from Table 2.2 (and the conclusions of Gallet 2007): its positive coefficient suggests a lower sensitivity to price changes in the short-run than in the long-run, when the enrollees have more time to adapt to a new pricing scheme and search for adequate substitutes.

Table 2.5 reports that the evidence on the importance of the *OLS* and *Control for endogeneity* variables is mixed across different model averaging approaches. The instability of the two coefficients is somewhat intuitive: studies using simple OLS rarely control for endogeneity; the correlation coefficient of these variables is -0.45. The direction of the effect of controlling for endogeneity that we identify is, however, not consistent with what is often found in the literature (Savoca 1990; Neill 2009): estimates that do not account for endogeneity are expected to show smaller effects since these estimates may capture the positive effects of price on the supply of education. Our results suggest that controlling for endogeneity understates the reported effects (although the corresponding posterior inclusion probability is less than 0.5, suggesting a very weak link). This finding is consistent with Gallet (2007), who reports that methods controlling for endogeneity generate more positive estimates than does OLS, and it suggests a potential problem with approaches to endogeneity

Table 2.5: Explaining heterogeneity in the estimates of the tuition-enrollment nexus

Response variable:	Bayesian model averaging			Frequentist check (OLS)			Frequentist model averaging		
Tuition PCC	Post. mean	Post. SD	PIP	Coef.	SE	p-value	Coef.	SE	p-value
Constant	0.002	NA	1.000	-0.149	0.051	0.004	0.005	0.003	0.086
Standard error	-0.650	0.439	0.758	-0.673	0.099	0.000	-0.712	0.252	0.005
Estimation characteristics									
Short-run effect	0.052	0.053	0.575	0.093	0.007	0.000	0.138	0.038	0.000
OLS	-0.097	0.067	0.742	-0.105	0.044	0.018	-0.016	0.052	0.766
Control for endogeneity	0.052	0.073	0.414				0.165	0.052	0.002
Design of the demand function									
Linear function	0.000	0.011	0.064				-0.072	0.052	0.172
Double-log function	-0.003	0.015	0.095				-0.068	0.046	0.134
Unemployment control	0.008	0.026	0.132				0.073	0.045	0.107
Income control	0.083	0.065	0.718	0.054	0.021	0.009	0.191	0.040	0.000
Data specifications									
Cross-sectional data	0.000	0.022	0.093				-0.085	0.050	0.091
Panel data	-0.112	0.073	0.784	-0.018	0.013	0.168	-0.238	0.055	0.000
Male candidates	-0.351	0.062	1.000	-0.227	0.073	0.002	-0.381	0.065	0.000
Female candidates	-0.007	0.040	0.072				-0.145	0.110	0.189
Private universities	-0.169	0.039	0.996	-0.077	0.003	0.000	-0.173	0.048	0.000
Public universities	0.001	0.012	0.060				0.011	0.038	0.767
USA	-0.095	0.086	0.643	-0.038	0.036	0.283	-0.196	0.060	0.001
Publication characteristics									
Publication year	-0.013	0.018	0.421				-0.016	0.013	0.212
Citations	0.043	0.010	1.000	0.033	0.003	0.000	0.053	0.010	0.000
Published study	-0.015	0.037	0.183				-0.054	0.052	0.301
Studies	43			43			43		
Observations	442			442			442		

Notes: SD = standard deviation. SE = standard error. PIP = posterior inclusion probability. Bayesian model averaging (BMA) employs priors suggested by Eicher *et al.* (2011). The frequentist check (OLS) includes the variables recognized by BMA to comprise the best model and is estimated using standard errors clustered at the study and country level. Frequentist model averaging (FMA) follows Mallow's averaging using the orthogonalization of covariate space suggested by Amini & Parmeter (2012). All variables are described in Table 2.4. Additional details on the BMA exercise can be found in the Section 2.B in Table 2.10 and Figure 2.8.

control employed in the literature. Alternatively, the finding might also imply that economists do not fully understand the demand for education.

Several techniques are commonly used that attempt to purge the effects of endogeneity from the demand equation, but difficulties for researchers often arise while doing so. In the presence of endogeneity, OLS should directly lead to biased and inconsistent estimates, although some researchers justify its utilization by the identification of the supply and demand side. First, with public financing and/or constant operating costs supported by revenues collected from new enrollees, the universities may supply more enrollments without raising tuition (King 1993). When applicants outstrip enrollment (which they often do), the tuition price in the face of excess demand does not clear the market for higher education. Researchers then assume the supply of admission at the market level of tuition to be infinitely elastic and all independent variables to explain only the demand side (justifying the utilization of OLS as in Mueller & Rockerbie 2005). Second, as Coelli (2009) states, one could explain the demand side in such a detailed manner that would obliterate the remaining endogeneity. For this purpose, Coelli (2009) uses a set of rich individual student and parental characteristics.

The perfect elasticity of the supply side is, however, rather academic (we can never rule out the correlation between the explanatory variable and the error term of the demand equation with certainty). Moreover, detailed micro-level information usually comes from surveys and is unavailable for most researchers, which makes the simultaneous equation model difficult to estimate. The literature treating endogeneity thus relies on instrumental variables instead. Instrumental variables are defined as tuition correlates that should not directly affect the demand for higher education. Choosing the appropriate instrument might be tricky: the instrument must be strong enough to provide a source of variation for the model but must still show an exogenous source of variation. Also the fixed-effect instrumental variable estimation with a weak instrument can easily lead to results that are no better than those of OLS.

The lagged level of the endogenous variable represents a popular choice for an instrument (Allen & Shen 1999). The tuition fee is typically further instrumented by an income level and unemployment rate (sometimes in lags, which reflect more realistic delayed adjustments in tuition), determinants of education costs (excess-tax on tuition, faculty salaries), and time dummies capturing other possible external influences. Two-stage and three-stage least squares typically appear in applications (Berger & Kostal 2002; Savoca 1990). The generalized-method-of-moments estimator (Arellano & Bond 1991) is another frequently chosen technique. Nevertheless, contrary to the recommendations of Roodman (2009), many studies do not report the number of instruments used in the analysis. In a nutshell, we believe that even the results of the studies that claim to successfully control for endogeneity cannot be automatically interpreted causally.

Design of the demand function: According to our results, the functional form of the demand function does not systematically affect the reported coefficients. This conclusion differs from the findings of Gallet (2007), who argues that the outputs of semi-log, linear, and Box-Cox functional forms are significantly different from the results produced by directly estimating the double-log demand function. Furthermore, the inclusion of the control variable for *unemployment* also does not seem to drive the estimated sensitivity of enrollment to tuition changes; the control for an individual's *income* group, however, significantly decreases the estimated sensitivity.

Data specifications: Leslie & Brinkman (1987) report that estimates produced from cross-sectional datasets and time-series datasets do not vary substantially, and our results support this conclusion. *Panel data*, which combine both cross-sectional and time information, however, lead to partial correlation coefficients that are 0.11 smaller, other things being equal. We also argue that *male students* exhibit a systematically larger (by 0.35) sensitivity to changes in tuition in comparison with the general population, which is in contrast to the results of Gallet (2007), who finds that gender-related characteristics fail to significantly affect the reported tuition elasticity. The results are, however, in line with those of Mueller & Rockerbie (2005), who find males to be more price sensitive than females. As an explanation, Mueller &

Rockerbie (2005) argue that since the rate of return to a university degree might be higher for a female than for a male, females are willing to spend more on tuition fees.

Some studies estimate the effect of tuition in public universities, while others consider private universities. We find that candidates applying to *private universities* display larger tuition elasticities. One interpretation of the different magnitudes of the price sensitivity is that the more or the better the substitutes are for a particular commodity, the higher the price sensitivity. In our case, students should be able to switch more easily to a substitute institution when private university tuition rises than when public university tuition rises, as the pool of substitutes for private institutions should be larger and also includes public universities (where costs are lower). These results, however, contradict those of Leslie & Brinkman (1987) and Hight (1975), who note that the average enrollee at a private university is rich and, thus, less price-elastic. Estimates for the US seem to be less negative than those for other countries. We would argue that given the extent of the US system of higher education, the pool of close substitutes might be larger in the US than in the rest of the world, where a single country hosts a smaller number of universities.

Publication characteristics: There are two results on publication characteristics that are consistent with the meta-analysis of Gallet (2007): the insignificance of publication year and publication status. *Publication year* may capture changes in methodological approaches; nevertheless, Table 2.5 indicates that the newer studies do not report systematically different results. Further, we show in Table 2.2 that the partial correlation coefficients reported in *Published studies* are arguably smaller than those in unpublished or unrefereed studies. The impact of other explanatory variables, however, erases this link; in fact, Table 2.5 suggests the publication status of a study does not matter for the magnitude of the estimates. More important is how much attention the paper attracts from readers, which is captured by the number of *citations*. Highly cited articles report less-sensitive estimates of the tuition-enrollment relationship.

Thus far, we have argued that the mean reported value of the tuition-enrollment partial correlation coefficient, −0.19 (shown in Table 2.2), is significantly exaggerated by the presence of publication bias. The effect absent publication bias, shown in Table 4.6, is close to zero. We have also seen that the effect is substantially influenced by data, method, and publication characteristics. To provide the reader with a 'rule-of-thumb' mean effect that controls for all these influences and potential biases, we construct a synthetic 'best-practice' study that employs our preferred choices with respect to all the sources of heterogeneity in the literature. The definition of best practice is subjective, but it is a useful check of the combined effect of various misspecifications and publication bias. Essentially, we create a weighted average of all estimates by estimating fitted values from the BMA and FMA specifications.

The ideal study that we imagine would be published in a refereed journal, highly cited, and recent; thus, we set all publication characteristics at the sample maxima (we censor, however, the number of citations at the 99% level due to the presence of outliers—although using the sample maximum would provide us with an even stronger result). We remove any sources of publication and endogeneity bias; thus, we set the *standard error* and *OLS* at the sample minima and the *control for endogeneity* at the sample maximum. We prefer the usage of broader datasets and favor the inclusion of controls for the economic environment, and thus, we set the panel dataset and controls for income and unemployment at the sample maxima. Moreover, we prefer the double-log functional form since it directly produces an elasticity and represents a measure with a clear interpretation that is independent of the current price level. We leave the remaining variables at their sample means.

The 'best-practice' estimation in Table 2.6 yields a partial coefficient of −0.037 with a 95% confidence interval of (−0.055; −0.019). The estimated standard errors are relatively small, and even with plausible changes to the definition of best practice (such as changing the design of the demand function), the results reported in Table 2.6 change only at the third decimal place. The best-practice estimation thus corroborates our previous assertions regarding the correlation between tuition and enrollment: in general, we observe higher price elasticities in the long-run, higher elasticities among individuals enrolled in private universities, and higher elasticities

Table 2.6: Best practice estimation yields a tuition-enrollment effect that is close to zero

	Bayesian model averaging			Frequentist model averaging		
	Mean	95% conf. int.		Mean	95% conf. int.	
Short-run effect	-0.010	-0.032	0.012	0.069	0.047	0.090
Long-run effect	-0.062	-0.081	-0.042	-0.070	-0.089	-0.050
Private universities	-0.167	-0.190	-0.144	-0.141	-0.164	-0.117
Public universities	0.003	-0.033	0.039	0.043	0.007	0.079
Male candidates	-0.361	-0.529	-0.192	-0.348	-0.517	-0.180
Female candidates	-0.017	-0.120	0.087	-0.112	-0.216	-0.008
All estimates	-0.037	-0.055	-0.019	-0.003	-0.021	0.015

Notes: The table presents mean estimates of the partial correlation coefficients implied by the Bayesian/frequentist model averaging and our definition of 'best practice.' Because BMA does not work with the concept of standard errors, the confidence intervals for BMA are approximate and constructed using the standard errors estimated by simple OLS with robust standard errors clustered at the study and country level.

among male students. The overall mean, however, is very close to zero.

2.5 Extensions and Robustness Checks

In this section, we pursue seven modifications of our baseline BMA model presented earlier. The results of these modifications are divided into two tables, one entitled 'robustness checks' (four specifications) and the other entitled 'extensions' (three specifications). Roughly speaking, robustness checks show the sensitivity of our main results to plausible changes in estimation strategy, while extensions provide new results. We start by discussing Table 2.7, which contains the robustness checks. The structure of the table is similar to what we have already seen in the previous section: the results shown include the posterior mean, posterior standard deviation, and posterior inclusion probability for each variable. The table is divided into four vertical panels, which mark different specifications.

The first specification in Table 2.7 uses a set of priors different from the one in our baseline model. In this robustness check, we follow Fernandez *et al.* (2001) and choose BRIC for g-prior; for model size, we use the beta-binomial random prior advocated by Ley & Steel (2009). Along with the plain vanilla UIP g-prior and uniform model prior used in the baseline model, the choice of priors in the robustness check constitutes arguably the most common combination of priors used in BMA. In our experience,

Table 2.7: Explaining heterogeneity in the estimates (robustness checks of Table 2.5)

Response variable:	Different BMA priors			Precision-weighted data			Elasticities			Spatial variation		
Tuition PCC	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP
Constant	0.002	NA	1.000	-0.017	NA	1.000	-0.011	NA	1.000	0.002	NA	1.000
Standard error	-0.614	0.444	0.728	-0.277	3.915	0.513	-0.860	0.226	0.996	-0.660	0.439	0.764
Estimation characteristics												
Short-run effect	0.058	0.054	0.617	0.023	0.031	0.417	0.374	0.209	0.879	0.051	0.053	0.562
OLS	-0.096	0.067	0.742	-0.001	0.013	0.075	0.004	0.058	0.110	-0.097	0.066	0.747
Control for endogeneity	0.054	0.073	0.417	0.012	0.024	0.262	0.256	0.256	0.577	0.051	0.072	0.406
Design of the demand function												
Linear function	-0.001	0.015	0.076	-0.105	0.038	1.000				0.001	0.011	0.062
Double log function	-0.004	0.016	0.101	-0.054	0.054	0.639				-0.004	0.015	0.094
Unemployment control	0.010	0.029	0.156	0.074	0.022	0.977	0.750	0.138	1.000	0.007	0.026	0.128
Income control	0.079	0.069	0.657	0.019	0.033	0.335	-0.754	0.133	1.000	0.082	0.065	0.713
Data specifications												
Cross-sectional data	-0.003	0.028	0.116	0.198	0.029	1.000	0.265	0.251	0.615	0.000	0.022	0.090
Panel data	-0.109	0.077	0.746	0.009	0.027	0.157	-0.606	0.126	1.000	-0.112	0.072	0.788
Male candidates	-0.349	0.063	1.000	-0.213	0.055	0.996	-0.012	0.113	0.074	-0.351	0.062	1.000
Female candidates	-0.010	0.049	0.091	-0.030	0.066	0.219	-0.009	0.107	0.071	-0.006	0.040	0.070
Private universities	-0.169	0.039	0.996	0.173	0.032	1.000	-0.886	0.105	1.000	-0.169	0.039	0.997
Public universities	0.001	0.012	0.065	0.167	0.028	1.000	-0.404	0.096	0.996	0.001	0.011	0.058
USA	-0.091	0.088	0.593	-0.183	0.031	1.000	-0.007	0.050	0.090	-0.093	0.086	0.628
Canada										0.000	0.002	0.054
Netherlands										0.001	0.005	0.084
Publication characteristics												
Publication year	-0.015	0.018	0.472	0.019	0.514	0.512	-0.001	0.011	0.088	-0.013	0.018	0.415
Citations	0.043	0.010	0.999	0.029	0.007	1.000	0.199	0.047	1.000	0.043	0.010	1.000
Published study	-0.015	0.038	0.190	-0.128	0.042	0.984	-0.022	0.095	0.116	-0.014	0.037	0.178
Studies	43			43			24			43		
Observations	442			442			224			442		

Notes: SD = standard deviation. PIP = posterior inclusion probability. *Different BMA prior* = benchmark from Table 2.5 estimated using BMA with model priors according to Fernandez *et al.* (2001) and Ley & Steel (2009). The corresponding visualization is represented by Figure 2.10, and the corresponding diagnostics of the BMA can be found in Table 2.11. *Precision-weighted data* = benchmark model with data weighted by the inverse of the standard error and estimated by BMA which uses model priors according to Eicher *et al.* (2011). The corresponding visualization is represented by Figure 2.11, and the corresponding diagnostics of the BMA can be found in Table 2.12. Different weighting schemes for meta-analysis are discussed in greater detail by Zigraiova & Havranek (2016, p. 28-30). *Elasticities* = only estimates coming from double-log functional forms are considered. *Spatial variation* = dummy variables for estimates coming from Canada and Netherlands are included in the model. All variables are described in Table 2.4.

the choice between UIP and BRIC rarely is material for BMA results. In contrast, the choice of model prior can have important effects.

The uniform model prior relies on the intuitive notion that each model, irrespective of the number of variables included, should have the same weight in results. Among all the possible models, the most common are those that contain the mean number of these variables (and the least common are those that include only a couple of them or almost all of them). In consequence, the uniform model prior puts more weight on models of mean size. An appealing alternative is to place the prior directly on model size, which is what we do here: we assume that ex ante, each model *size* has the same probability. In this application, the effect of this change on results is minimal. The estimated coefficients change little; we still obtain evidence of significant publication bias, and our discussion of the impact of individual variables would not change at all. This is an encouraging finding of robustness to a choice in priors, strengthened further by the fact that frequentist model averaging (which relies on a completely different econometric philosophy), presented in the previous section, delivers similar results as well.

In the second panel of Table 2.7, we use weights proportional to the reported precision of the estimates. These are appealing weights because they address the inherent heteroskedasticity problem in meta-analysis, and intuitively, it makes sense to give more weight to more precise results. Indeed, throughout the analysis, we have treated the estimates of the tuition-enrollment nexus as data points, although the estimates have uncertainty attached to them. The appendix to Havranek *et al.* (2017) shows technically how this uncertainty affects the results of Bayesian model averaging. In sum, our posterior inclusion probabilities might be exaggerated. This problem can be addressed in two ways: first, by estimating OLS or frequentist model averaging with robust standard errors (as we did in the previous section), and second, by estimating a BMA model with weights proportional to precision. This is an intuitive solution that also follows the literature on estimated dependent variable models, but such a precision-weighted BMA exercise lacks rigorous Bayesian foundations, as is also discussed by Havranek *et al.* (2017). Nevertheless, we proceed with the robustness check.

The precision-weighted BMA is more difficult to interpret because of increased collinearity. This issue is especially serious for variables that show little variation within studies; by using weights proportional to precision, we generate artificial correlation between these variables. In consequence, it is hard to comment on individual variables—with the exception of the standard error itself. We see from Table 2.7 that while the corresponding coefficient retains its negative sign, its magnitude drops. Additionally, the posterior inclusion probability decreases, although it still surpasses the 0.5 threshold: the underlying model is likely to include a variable that corrects for publication bias. Perhaps the most useful comparison between this robustness check and our baseline model involves computing the best-practice estimate: after all, that is the main output of our analysis. Because such an estimate uses the specification as a whole, it is not affected by collinearity problems. Here, we get a best-practice estimate of -0.04, which is very close to our baseline result and consistent with the notion that after correcting for publication bias and misspecifications, there is little evidence for any strong tuition-enrollment nexus.

The third panel of Table 2.7 focuses on the subset of estimates that are computed using the log-log specification. Consequentially, they can be interpreted as elasticities and allow for more direct discussion of the underlying economic effect of tuition fees. This is obviously a useful robustness check, but we have to exclude approximately 50% of the dataset for which we are unable to reconstruct elasticities (very often, sample means for the data are not reported in the literature, so we cannot approximate elasticities for linear and other estimates). It is also clear that we have to omit from the BMA exercise variables that reflect the functional form used in primary studies, because now we only focus on log-log specifications. The estimated model thus changes somewhat, but evidence for publication bias is even stronger than in our baseline model: the posterior inclusion probability for this variable reaches 0.996. The implied best-practice estimate is, once again, statistically indistinguishable from zero.

In the last panel of Table 2.7, we add two more dummy variables reflecting spatial variation in the estimates of the tuition-enrollment nexus. A large majority of the estimates in our sample was derived from US data, but the next two countries with the largest number of estimates are Canada and the Netherlands (in total, our sample covers 21 countries, but for many of them, we only have a few estimates, so it is unfeasible to include dummy variables for all of them). Controlling for potential systematic differences in estimates for these two countries does not change our results, and the two dummy variables have low posterior inclusion probabilities (less than 0.1).

Now, we turn our attention to extensions of the baseline model. Sometimes it is not easy to draw a line between robustness checks and extensions (for example, the additional country dummies in the last panel of Table 2.7 might also be considered an extension), but the distinction is useful for presentation purposes. The core of the extensions presented in Table 2.9 consists of the addition of new potentially useful variables into the model. The reason that we do not consider these variables in the baseline model is the same reason that we do not, in the baseline, use weights that are proportional to precision: the inclusion of these new variables greatly increases collinearity, so we find it more useful to present them in a separate exercise. The new variables are first presented in Table 2.8, where we show sample means for the respective groups of estimates.

First, we include a dummy variable *Individual level*, which equals one for estimates that are computed using student-level data, as opposed to more aggregated data available at the university level. Individual-level estimates are relatively rare in the literature and, on average, seem to show larger partial correlation coefficients for the tuition-enrollment nexus. Next, we account for the type of study program under examination: undergraduate, graduate, and MBA. We also include an additional dummy variable that equals one if the study focuses on freshman students. We observe larger partial correlation coefficients for graduates and MBAs (which makes good sense, because these products still have a special status, and students can easily choose to work instead, thereby increasing the price elasticity), but it is important to note that for these groups, we have relatively few observations.

Table 2.8: Partial correlation coefficients for different subsets of data (additional variables)

		Unweighted			Weighted		
	Observations	Mean	95% conf. int.		Mean	95% conf. int.	
Heterogeneity							
Individual level	85	-0.294	-0.352	-0.236	-0.234	-0.299	-0.168
University level	357	-0.141	-0.169	-0.114	-0.174	-0.206	-0.143
Undergraduate enrollees	288	-0.118	-0.148	-0.088	-0.169	-0.206	-0.133
Graduate enrollees	34	-0.420	-0.523	-0.316	-0.342	-0.421	-0.264
Freshman enrollees	132	-0.124	-0.163	-0.084	-0.190	-0.243	-0.138
MBA program	8	-0.324	-0.462	-0.186	-0.328	-0.458	-0.199
Spatial variation							
Canada	51	-0.364	-0.430	-0.299	-0.200	-0.270	-0.130
Netherlands	24	-0.093	-0.139	-0.046	-0.063	-0.105	-0.022
All estimates	442	-0.171	-0.196	-0.145	-0.186	-0.214	-0.158

Notes: The table reports mean values of the partial correlation coefficients for different subsets of data. The exact definitions of the variables are available in Table 2.4. Weighted = estimates that are weighted by the inverse of the number of estimates per study.

In the first panel of Table 2.9, we explore the consequences of adding these new variables into our baseline model. In addition to variables listed in Table 2.8, we also include the square root of the number of observations used in the primary study: because partial correlation coefficients are a function of the number of observations, this variable may be important by definition. As we have noted earlier, collinearity increases above acceptable limits (many variables now have variance-inflation factors above 10), so the interpretation of the signs and magnitudes of individual coefficients should be taken with a grain of salt. That being said, the parameter corresponding to the magnitude of publication bias retains its negative sign but loses much of its importance (the posterior mean is now only -0.16 and posterior model probability is 0.28). Concerning the additional variables, two findings arise that also survive a later robustness check: first, ceteris paribus, studies using student-level data tend to find larger partial correlation coefficients than studies using university level data; second, MBA programs are also associated with a stronger tuition-enrollment nexus. Undergraduate and graduate study programs seem to display similar price elasticities, and the correlation between the number of observations and reported partial correlation is small (and disappears in a robustness check).

Table 2.9: Explaining heterogeneity in the estimates (extensions of Table 2.5)

Response variable:	Additional variables			Time variation			Combined		
Tuition PCC	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP	Post. mean	Post. SD	PIP
Constant	0.001	NA	1.000	-0.001	NA	1.000	-0.004	NA	1.000
Standard error	-0.160	0.298	0.277	-1.089	0.294	0.989	-1.392	0.248	0.999
Estimation characteristics									
Short-run effect	0.000	0.009	0.048	0.169	0.045	0.993	0.001	0.009	0.058
OLS	-0.005	0.020	0.100	-0.055	0.075	0.431	-0.074	0.061	0.669
Control for endogeneity	0.064	0.059	0.613	0.182	0.092	0.899	0.016	0.037	0.201
Design of the demand function									
Linear function	0.060	0.070	0.485	-0.014	0.033	0.212	0.005	0.025	0.077
Double log function	-0.047	0.055	0.484	0.000	0.012	0.062	-0.134	0.040	0.967
Unemployment control	0.010	0.028	0.154	0.009	0.027	0.156	0.011	0.031	0.162
Income control	0.001	0.010	0.058	0.209	0.041	1.000	0.001	0.008	0.046
Data specifications									
Cross-sectional data	0.105	0.055	0.861	-0.084	0.089	0.534	-0.001	0.011	0.051
Panel data	-0.010	0.032	0.146	0.021	0.055	0.177	0.293	0.063	0.999
Male candidates	-0.431	0.067	1.000	-0.372	0.062	1.000	-0.451	0.061	1.000
Female candidates	-0.037	0.092	0.187	-0.009	0.041	0.083	-0.008	0.040	0.079
Private universities	-0.174	0.039	0.999	-0.215	0.057	0.989	-0.252	0.051	1.000
Public universities	0.000	0.009	0.044	0.052	0.057	0.533	0.044	0.053	0.484
USA	-0.353	0.063	1.000	0.000	0.014	0.053	-0.173	0.069	0.941
Publication characteristics									
Publication year	-0.001	0.005	0.092	-0.046	0.013	0.982	-0.002	0.008	0.105
Citations	0.040	0.010	0.998	0.071	0.012	1.000	0.083	0.010	1.000
Published study	0.036	0.062	0.324	-0.004	0.020	0.081	0.183	0.048	0.994
Additional variables									
Individual level	-0.441	0.068	1.000				-0.438	0.056	1.000
Undergraduate enrollees	0.009	0.026	0.157				0.000	0.009	0.052
Graduate enrollees	-0.049	0.112	0.200				-0.026	0.093	0.111
Freshmen enrollees	0.097	0.058	0.819				0.025	0.040	0.343
MBA program	-0.295	0.150	0.839				-0.415	0.137	0.942
Sqrt(observations)	0.004	0.002	0.851				0.000	0.001	0.097
Time variation									
Publication year x SE				0.009	0.048	0.118	-0.003	0.023	0.057
Publication year x OLS				-0.006	0.010	0.304	0.000	0.002	0.053
Publication year x endogeneity				-0.004	0.010	0.217	0.002	0.006	0.186
Publication year x cross-section				-0.015	0.017	0.513	0.000	0.002	0.047
Publication year x panel				-0.063	0.012	1.000	-0.087	0.012	1.000
Studies	43			43			43		
Observations	442			442			442		

Notes: SD = standard deviation. PIP = posterior inclusion probability. All models are estimated using Bayesian model averaging employing priors suggested by Eicher *et al.* (2011).

The second panel of Table 2.9 shows how the effect of selected variables changes in time. For this extension we choose five important variables: standard error (reflecting publication bias), the use of OLS, control for endogeneity, use of cross-sectional data, and use of panel data. In our baseline BMA model, we find no systematic influence of the publication year of studies on the reported results. In contrast, we find evidence of systematic effects related to the use of OLS and panel data. In this extension, we find no evidence that the magnitude of publication bias would change in time (the interaction of the reported standard error and the publication year of the study has a posterior inclusion probability of about 0.12). A large posterior inclusion probability (1) is reported for the interaction between the use of panel data and publication year. The interaction is negative, which means that the negative effect of the use of panel data on the reported results identified in our baseline model has been strengthening in recent years. In the third column of the table, we merge both extensions, which provides a simple robustness check. Overall, these two extensions suggest that data aggregation systematically affects results and that there is limited evidence for the literature converging to a consensus (which could be reflected, for example, by a trend in publication bias).

2.6 Concluding Remarks

In this paper, we conduct a quantitative synthesis of 442 estimates of the relationship between tuition fees and the demand for higher education reported in 43 studies. Our contribution on top of the previous meta-analysis by Gallet (2007) is twofold: first, we include a formal treatment of publication bias, and second, we include a treatment of model uncertainty using model averaging methods when searching for the determinants of the underlying effect. The literature shows substantial publication selection against positive estimates, suggesting that many researchers use the sign of the estimated effect as a specification test (education is unlikely to be a Giffen good). The mean effect beyond publication bias is close to zero. When we attribute greater weight to the more reliable estimates (published in respected journals and derived using appropriate methodology), we obtain a similarly small mean estimate of the tuition-enrollment nexus.

We also find evidence for systematic dependencies between the estimated effects and data, methodological, and publication characteristics. Male students display larger tuition elasticities, as do students at private universities. Previous research has yielded mixed results on both of these relationships. Our findings concerning male students are consistent with those of Mueller & Rockerbie (2005), who argue that because female students tend to have a higher rate of return from university education, they are willing to spend more on tuition fees. Concerning private universities, it might be easier for their students to find substitutes if tuition increases; for public university students, a large portion of the market (most private universities) is already unaffordable. Next, we find that highly cited studies tend to report little correlation between enrollment and tuition, although the direction of causality is unclear. Our results also suggest that the reported relationship is larger for US students and when panel data are used, while it is lower when income is controlled for and in the short run. Moreover, it is remarkable that the correlation between tuition and enrollment has been stable in time over the last 50 years.

Two qualifications of our analysis are in order. First, while we would prefer to work with elasticities, many studies estimate the relationship between tuition fees and enrollment using approaches other than the log-log specification. We already have to exclude a significant portion of studies because they do not report standard errors, t-statistics, or confidence intervals for their results, thus making it impossible for us to test the presence of publication bias. Restricting our dataset to log-log specifications would drastically reduce the number of degrees of freedom available for our analysis. While it is possible to recompute some of the other coefficients to elasticities evaluated at the sample mean, many studies do not report the statistics necessary for this computation. Therefore, we choose to work with partial correlation coefficients, which can be computed easily from all the studies, and we include an analysis of the sub-sample of elasticities only as an extension. Since our main result indicates negligible partial correlation absent publication bias, it also directly translates into a finding of a zero mean elasticity of demand for higher education to tuition fees. Second, the results of a meta-analysis are clearly conditional on the quality of the previous studies included in the sample. For instance, if all studies in the literature share a common misspecification that biases their results toward zero, we are unable to control for such a misspecification, and our results are thus also biased. Therefore, the correct interpretation of our analysis is that, judging from the available empirical research, our best guess concerning the effect of tuition on enrollment is close to zero.

References

Agarwal, V. B. & D. R. Winkler (1985): "Foreign Demand for United States Higher Education: A Study of Developing Countries in the Eastern Hemisphere." *Economic Development and Cultural Change* **33(3)**: pp. 623–44.

Alexander, E. & D. Frey (1984): "An econometric estimate of the demand for MBA enrollment." *Economics of Education Review* **3(2)**: pp. 97–103.

Allen, R. & J. Shen (1999): "Some new evidence of the character of competition among higher education institutions." *Economics of Education Review* **18(4)**: pp. 465–470.

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

Arellano, M. & S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equation." *The Review of Economic Studies* **58(2)**: pp. 277–297.

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

Babecky, J. & T. Havranek (2014): "Structural reforms and growth in transition." *The Economics of Transition* **22(1)**: pp. 13–42.

Benos, N. & S. Zotou (2014): "Education and Economic Growth: A Meta-Regression Analysis." *World Development* **64(C)**: pp. 669–689.

Berger, M. & T. Kostal (2002): "Financial resources, regulation, and enrollment in US public higher education." *Economics of Education Review* **21(2)**: pp. 101–110.

Bezmen, T. & C. Depken (1998): "School characteristics and the demand for college." *Economics of Education Review* **17(2)**: pp. 205–210.

Bruckmeier, K., G. Fischer, & B. Wigger (2013): "The willingness to pay for higher education: does the type of fee matter?" *Applied Economics Letters* **20(13)**: pp. 1279–1282.

Bruckmeier, K. & B. U. Wigger (2014): "The effects of tuition fees on transition from high school to university in Germany." *Economics of Education Review* **41**: pp. 14 – 23.

Buss, C., J. Parker, & J. Rivenburg (2004): "Cost, quality and enrollment demand at liberal arts colleges." *Economics of Education Review* **23(1)**: pp. 57–65.

Cameron, A. C., J. B. Gelbach, & D. L. Miller (2011): "Robust Inference With Multiway Clustering." *Journal of Business & Economic Statistics* **29(2)**: pp. 238–249.

Campbell, R. & B. N. Siegel (1967): "The demand for higher education in the United States, 1919-1964." *The American Economic Review* **57(3)**: pp. 482–294.

Campos, J., N. Ericsson, & D. Hendry (2005): *General-to-Specific Modeling*, chapter General-to-Specific Modeling: An Overview and Selected Bibliography. Edward Elgar Publishing.

Canton, E. & F. de Jong (2002): "The Demand for Higher Education in the Netherlands, 1950-'99." *Economics of Education Review* **24(6)**: pp. 651–663.

Chen, S. (2016): "College Costs, Freshman Enrollment, and Student Quality: Evidence from an Urban State University." *International Journal of Information and Education Technology* **6(9)**: pp. 664–671.

CHESLOCK, J. (2001): "The Determinants of an Institution's Transfer Student Enrollment." *Working paper 20/2001*, Center for the Study of Higher Education at the University of Arizona.

CHISHOLM, M. & B. COHEN (1982): "A Review and Introduction to Higher Education Price Response Studies." *Working paper series*, Boulder, CO: National Center for Higher Education Management Systems, http://files.eric.ed.gov/fulltext/ED246805.pdf.

CHRESSANTHIS, G. (1986): "The impacts of tuition rate changes on college undergraduate headcounts and credit hours over time: A case study." *Economics of Education Review* **5(2)**: pp. 205–217.

CHRISTOFIDES, L., J. CIRELLO, & H. M. (2001): "Family Income and Postsecondary Education In Canada." *Canadian Journal of Economics/Revue canadienne d'économique* **31(1)**: pp. 177–208.

COELLI, M. (2009): "Tuition fees and equality of university enrolment." *Canadian Journal of Economics/Revue canadienne d'économique* **42(3)**: pp. 1072–1099.

COHEN, J. (1988): *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

CORMAN, H. & P. DAVIDSON (1984): "Economic Aspects of Post-secondary Schooling Decisions." *Economics of Educonon Review* **3(2)**: pp. 131–139.

CRAFT, R., J. BAKER, B. MYERS, & A. HARRAF (2012): "Tuition revenues and enrollment demand: The case of southern Utah university." *Report 124/2012*, Association for Institutional research, Tallahassee, Florida.

DEARDEN, L., E. FITZSIMONS, & G. WYNESS (2011): "The Impact of Tuition Fees and Support on University Participation in the UK." *Working paper 126/2011*, Centre for the Economics of Education, London School of Economics.

DJANKOV, S. & P. MURRELL (2002): "Enterprise Restructuring in Transition: A Quantitative Survey." *Journal of Economic Literature* **40(3)**: pp. 739–792.

DOUCOULIAGOS, C. (1995): "Worker Participation and Productivity in Labor-Managed and Participatory Capitalist Firms: A Meta-Analysis." *ILR Review*

49(1): pp. 58–77.

Doucouliagos, C. (2011): "How Large is Large? Preliminary and relative guidelines for interpreting partial correlations in economics." *Economics Series 2011/5*, Deakin University, Faculty of Business and Law, School of Accounting, Economics and Finance.

Doucouliagos, C. & P. Laroche (2003): "What Do Unions Do to Productivity? A Meta-Analysis." *Industrial Relations* **42(4)**: pp. 650–691.

Doucouliagos, C. & T. D. Stanley (2013): "Are all economic facts greatly exaggerated? Theory competition and selectivity." *Journal of Economic Surveys* **27(2)**: pp. 316–339.

Doyle, C. & J. Cicarelli (1980): "The demand for higher education: a disaggregate approach." *The American Economist* **24(2)**: pp. 53–55.

Dwenger, N., J. Storck, & K. Wrohlich (2012): "Do tuition fees affect the mobility of university applicants? evidence from a natural experiment." *Economics of Education Review* **31(1)**: pp. 155–167.

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

Elliott, C. & K. Soo (2013): "The international market for MBA qualifications: The relationship between tuition fees and applications." *Economics of Education Review* **34**: pp. 162–174.

Espey, M. (1998): "Gasoline demand revisited: an international meta-analysis of elasticities." *Energy Economics* **20(3)**: pp. 273–295.

Feldkircher, M. & S. Zeugner (2009): "Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging." *IMF Working Papers 09/202/2009*, International Monetary Fund.

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

FLEURY, N. & F. GILLES (2015): "A meta-regression analysis on intergenerational transmission of education: publication bias and genuine empirical effect." *TEPP Working Paper 02/2015*, TEPP.

FUNK, H. J. (1972): "Price Elasticity of Demand for Education at a Private University." *Journal of Educational Research* **66(130-134)**.

GALLET, C. (2007): "A Comparative Analysis of the Demand for Higher Education: Results from a Meta-Analysis of Elasticities." *Economics Bulletin* **9(7)**: pp. 1–14.

GALPER, H. & R. M. DUNN (1969): "A Short-Run Demand Function for Higher Education in the United States." *Journal of Political Economy* **77(5)**: pp. 765–777.

GRUBB, W. (1988): "Vocationalizing Higher Education: the Causes of Enrollment and Completion in Public Two-Year Colleges, 1970-1980." *Economics of Education Review* **7(3)**: pp. 301–319.

HANSEN, B. (2007): "Least Squares Model Averaging." *Econometrica* **75(4)**: pp. 1175–1189.

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

HAVRANEK, T., D. HERMAN, & Z. IRSOVA (2018a): "Does Daylight Saving Save Electricity? A Meta-Analysis." *Energy Journal* **39(2)**: pp. 35–61.

HAVRANEK, T., R. HORVATH, Z. IRSOVA, & M. RUSNAK (2015a): "Cross-Country Heterogeneity in Intertemporal Substitution." *Journal of International Economics* **96(1)**: pp. 100–118.

HAVRANEK, T., R. HORVATH, & A. ZEYNALOV (2016): "Natural Resources and Economic Growth: A Meta-Analysis." *World Development* **88**: pp. 134 – 151.

HAVRANEK, T. & Z. IRSOVA (2011): "Estimating Vertical Spillovers from FDI: Why Results Vary and What the True Effect Is." *Journal of International Economics* **85(2)**: pp. 234–244.

Havranek, T. & Z. Irsova (2012): "Survey Article: Publication Bias in the Literature on Foreign Direct Investment Spillovers." *Journal of Development Studies* **48(10)**: pp. 1375–1396.

Havranek, T. & Z. Irsova (2017): "Do Borders Really Slash Trade? A Meta-Analysis." *IMF Economic Review* **65(2)**: pp. 365–396.

Havranek, T., Z. Irsova, & K. Janda (2012): "Demand for Gasoline is More Price-Inelastic than Commonly Thought." *Energy Economics* **34(1)**: pp. 201–207.

Havranek, T., Z. Irsova, K. Janda, & D. Zilberman (2015b): "Selective Reporting and the Social Cost of Carbon." *Energy Economics* **51(C)**: pp. 394–406.

Havranek, T., Z. Irsova, & T. Vlach (2018b): "Measuring the Income Elasticity of Water Demand: The Importance of Publication and Endogeneity Biases." *Land Economics* **94(2)**: pp. 259–283.

Havranek, T. & O. Kokes (2015): "Income Elasticity of Gasoline Demand: A Meta-Analysis." *Energy Economics* **47(C)**: pp. 77–86.

Havranek, T. & M. Rusnak (2013): "Transmission Lags of Monetary Policy: A Meta-Analysis." *International Journal of Central Banking* **9(4)**: pp. 39–76.

Havranek, T., M. Rusnak, & A. Sokolova (2017): "Habit Formation in Consumption: A Meta-Analysis." *European Economic Review* **95(1)**: pp. 142–167.

Heller, D. E. (1997): "Student Price Response in Higher Education: An Update to Leslie and Brinkman." *The Journal of Higher Education* **68(6)**: pp. 624–659.

Hemelt, S. & D. Marcotte (2011): "The Impact of Tuition Increases on Enrollment at Public Colleges and Universities." *Educational Evaluation and Policy Analysis* **33(4)**: pp. 435–357.

Hight, J. (1975): "The demand for higher education in the US 1927-72: The public and private institutions." *The Journal of Human Resources* **10(4)**: pp. 512–520.

Hoenack, S. & D. Pierro (1990): "An econometric model of a public university's income and enrollments." *Journal of Economic Behavior & Organization* **14(3)**: pp. 403–423.

Hoenack, S. & W. Weiler (1975): "Cost-related tuition policies and university enrollments." *Journal of Human Resources* **10(3)**: pp. 332–360.

Hoenack, S. A. (1971): "The efficient allocation of subsidies to college students." *American Economic Review* **61(3)**: pp. 302–11.

Hopkins, T. (1974): "Higher education enrollment demand." *Economic Inquiry* **12**: pp. 53–65.

Hsing, Y. & H. Chang (1996): "Testing increasing sensitivity of enrollment at private institutions to tuition and other costs." *The American Economist* **40(1)**: pp. 40–45.

Huijsman, R., T. Kloek, D. Kodde, & J. Ritzen (1986): "An empirical analysis of college enrollment in the Netherlands." *De Economist* **134(2)**: pp. 181–190.

Hübner, M. (2012): "Do tuition fees affect enrollment behavior? evidence from a 'natural experiment' in germany." *Economics of Education Review* **31(6)**: pp. 949–960.

Ioannidis, J. P. A., T. D. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *Economic Journal* **127**: pp. F236–F265.

Irsova, Z. & T. Havranek (2013): "Determinants of Horizontal Spillovers from FDI: Evidence from a Large Meta-Analysis." *World Development* **42(C)**: pp. 1–15.

Jackson, G. A. & G. B. Weathersby (1975): "Individual demand for higher education: A review and analysis of recent empirical studies." *The Journal of Higher Education* **46(6)**: pp. 623–651.

Jeffreys, H. (1961): *Theory of probability.* Oxford Classic Texts in the Physical Sciences. Oxford University Press, Oxford, 3 edition.

Kane, T. (2007): "Evaluating the impact of the DC tuition assistance grant program." *Journal of Human Resources* **42(3)**: pp. 555–582.

King, J. (1993): "The demand for higher education in Puerto Rico." *Economics of Education Review* **12(3)**: pp. 257–265.

Knudsen, O. & P. Servelle (1978): "The demand for higher education at private institutions of moderate selectivity." *The American Economist* **22(2)**: pp. 30–34.

Koshal, R., L. Gallaway, & R. Akkihal (1976): "Determinants of male and female higher education in the United States." *Social Indicators Research* **3(1)**: pp. 111–121.

Leslie, L. L. & P. T. Brinkman (1987): "Student Price Response in Higher Education: The Student Demand Studies." *The Journal of Higher Education* **58(2)**: pp. 181–204.

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

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

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

Mattila, J. P. (1982): "Determinants of Male School Enrollments: A Time-Series Analysis." *The Review of Economics and Statistics* **64(2)**: pp. 242–51.

McDuff, D. (2007): "Quality, tuition, and applications to in-state public colleges." *Economics of Education Review* **26(4)**: pp. 433–449.

McPherson, M. & M. Schapiro (1991): "Does student aid affect college enrollment? New evidence on a persistent controversy." *The American Economic Review* **81(1)**: pp. 309–318.

McPherson, M. S. (1978): "The Demand for Higher Education." In D. W. Breneman & C. E. J. Finn (editors), "Public Policy and Private Higher Education," pp. 143–196. Washington, D.C.: The Brookings Institutions.

Mueller, R. & D. Rockerbie (2005): "Determining demand for university education in Ontario by type of student." *Economics of Education Review* **24(4)**: pp.

469–483.

Murphy, R. & G. Trandel (1994): "The relation between a university's football record and the size of its applicant pool." *Economics of Education Review* **13(3)**: pp. 265–270.

Neill, C. (2009): "Tuition fees and the demand for university places." *Economics of Education Review* **28(5)**: pp. 561–570.

Noorbakhsh, A. & D. Culp (2002): "The demand for higher education: Pennsylvania's nonresident tuition experience." *Economics of Education Review* **21(3)**: pp. 277–286.

Ordovensky, J. (1995): "Effects of institutional attributes on enrollment choice: Implications for postsecondary vocational education." *Economics of Education Review* **12(4)**: pp. 311–324.

Ostheimer, R. H. (1953): *Student Charges and Financing Higher Education.* New York: Columbia University Press.

Parker, J. & J. Summers (1993): "Tuition and enrollment yield at selective liberal arts colleges." *Economics of Education Review* **12(4)**: pp. 311–324.

Paulsen, M. & T. Pogue (1988): "Higher education enrollment: The interaction of labor market conditions, curriculum, and selectivity." *Economics of Education Review* **7(3)**: pp. 275–290.

Quigley, J. & D. Rubinfeld (1993): "Public Choices in Public Higher Education." In C. T. Clotfelter & M. Rothschild (editors), "Studies of Supply and Demand in Higher Education," pp. 243–284. National Bureau of Economic Research, University of Chicago Press.

Quinn, R. & J. Price (1998): "The demand for medical education: An augmented human capital approach." *Economics of Education Review* **17(3)**: pp. 337–347.

Roodman, D. (2009): "A Note on the Theme of Too Many Instruments." *Oxford Bulletin of Economics and Statistics* **71(1)**: pp. 135–158.

Rusnak, M., T. Havranek, & R. Horvath (2013): "How to Solve the Price Puzzle?

A Meta-Analysis." *Journal of Money, Credit and Banking* **45(1)**: pp. 37–70.

Savoca, E. (1990): "Another look at the demand for higher education: Measuring the price sensitivity of the decision to apply to college." *Economics of Education Review* **9(2)**: pp. 123–134.

Shin, J. & S. Milton (2008): "Student response to tuition increase by academic majors: Empirical grounds for a cost-related tuition policy." *Higher Education* **55(6)**: pp. 719–734.

Stanley, T. & H. Doucouliagos (2010): "Picture This: A Simple Graph That Reveals Much Ado About Research." *Journal of Economic Surveys* **24(1)**: pp. 170–191.

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

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

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

Stanley, T. D. & S. B. Jarrell (1989): "Meta-Regression Analysis: A Quantitative Method of Literature Surveys." *Journal of Economic Surveys* **3(2)**: pp. 161–170.

Suloc, J. (1982): "The demand for community college education." *Economics of Education Review* **2(4)**: pp. 351–361.

Tannen, M. (1978): "The investment motive for attending college." *Industrial and Labor Relations Review* **31(4)**: pp. 489–497.

Toutkoushian, R. & P. Hollis (1998): "Using panel data to examine legislative demand for higher education." *Education Economics* **6(2)**: pp. 141–157.

Tuckman, H. (1970): "Determinants of college student migration." *Southern Economic Journal* **37(2)**: pp. 184–189.

Valickova, P., T. Havranek, & R. Horvath (2015): "Financial Development And

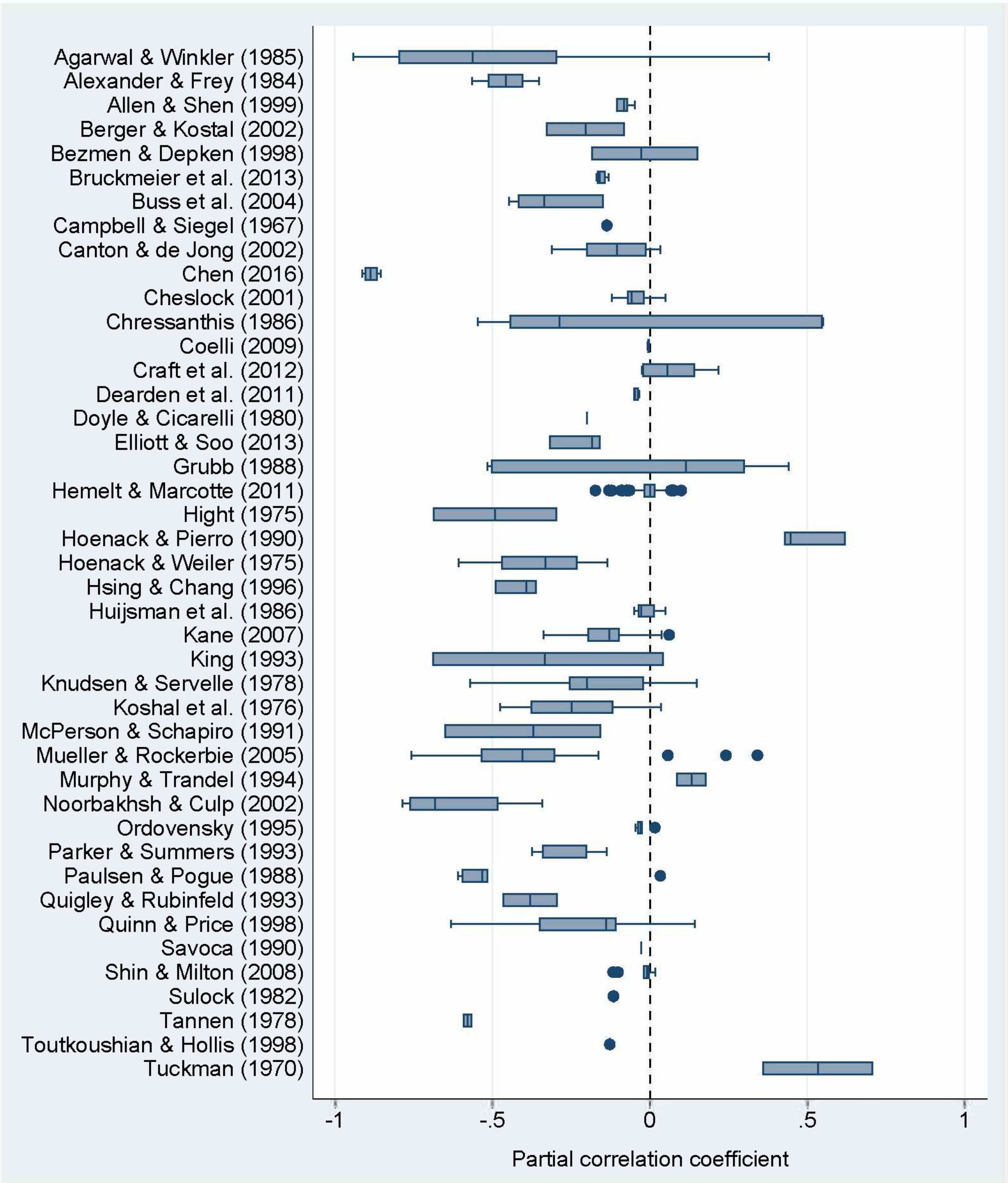
Economic Growth: A Meta-Analysis." *Journal of Economic Surveys* **29(3)**: pp. 506–526.

Weinschrott, D. J. (1977): "Demand for Higher Education in the U.S.: A Critical Review of the Empirical Literature." *Report R-2195-LE*, Santa Monica, CA: RAND Corporation.

Zigraiova, D. & T. Havranek (2016): "Bank competition and financial stability: Much ado about nothing?" *Journal of Economic Surveys* **30(5)**: pp. 944–981.

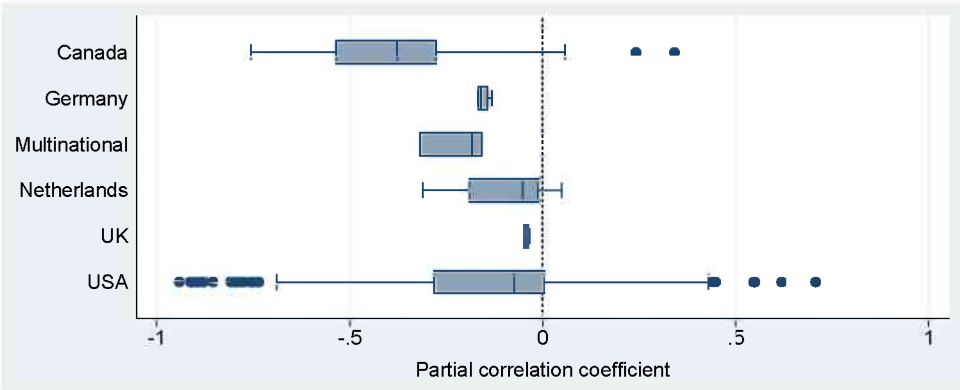
2.A Supplementary Statistics and Diagnostics of BMA

Figure 2.5: Estimates of the tuition-enrollment nexus vary within and across studies



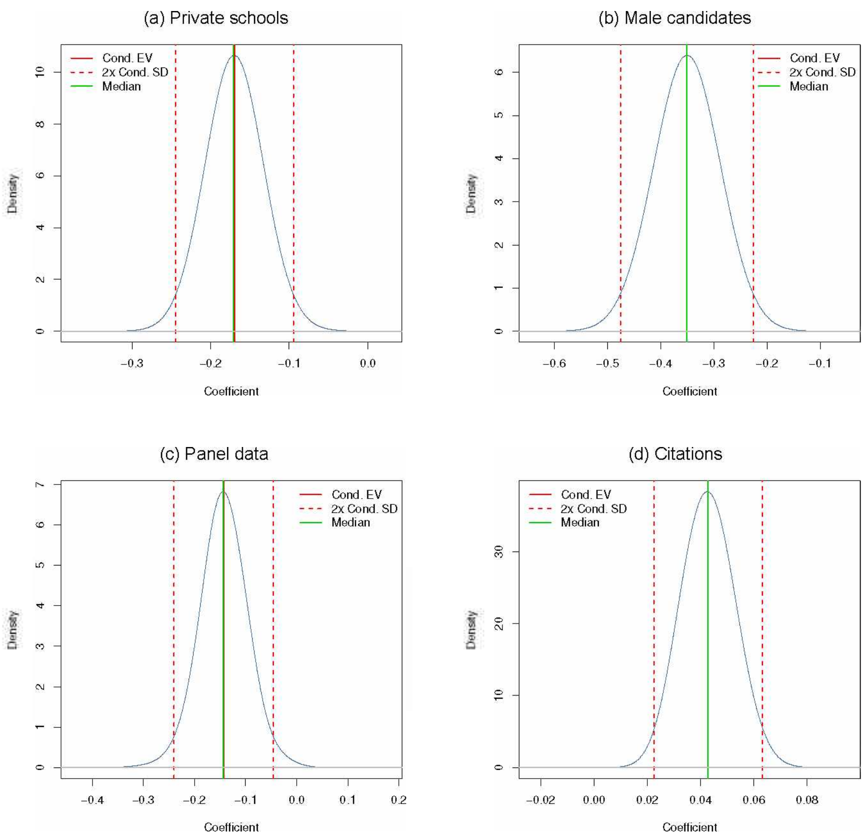
Notes: The figure shows a box plot of the partial correlation coefficients capturing the relationship between tuition and the demand for higher education reported in individual studies.

Figure 2.6: Estimates of the elasticity vary across different countries



Notes: The figure shows a box plot of the partial correlation coefficients capturing the relationship between tuition and the demand for higher education reported for individual countries.

Figure 2.7: Posterior coefficient distributions for the most important characteristics



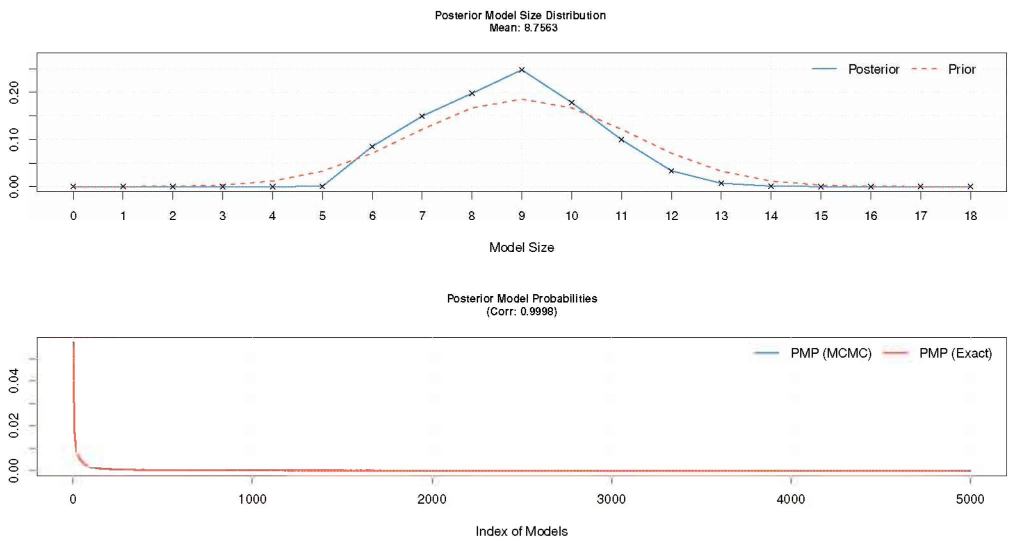
Notes: The figure depicts the densities of the regression parameters from Table 2.5 with the highest posterior inclusion probabilities.

Table 2.10: Summary of main BMA estimation

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.7563	$2 \cdot 10^6$	$1 \cdot 10^5$	5.495486 mins	578,591
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
262,144	22.10%	100%	0.9988	442
Model prior	*g-prior*	*Shrinkage-stats*		
Uniform	UIP	Av = 0.9977		

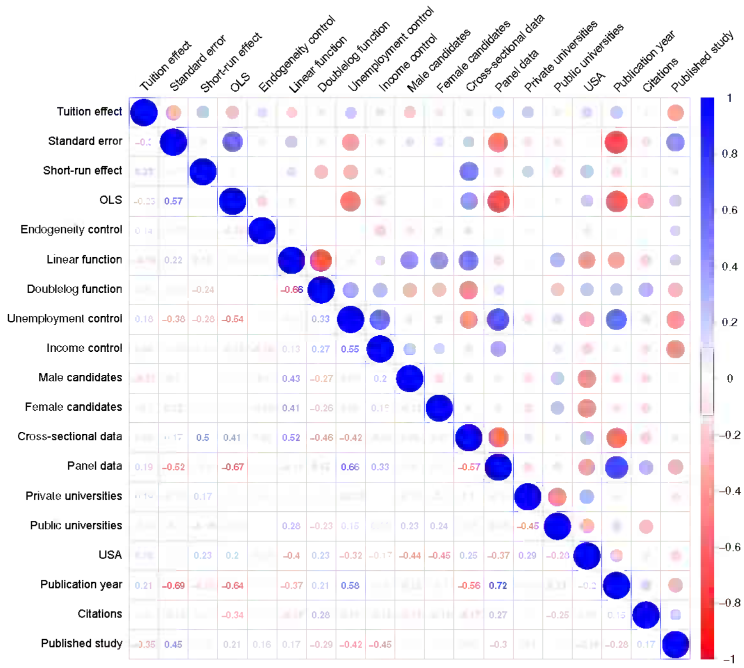
Notes: We employ the priors suggested by Eicher *et al.* (2011), who recommend using the uniform model prior (each model has the same prior probability) and the unit information prior (the prior provides the same amount of information as one observation in the data). The results of this BMA exercise are reported in Table 2.5.

Figure 2.8: Model size and convergence of main BMA estimation



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

Figure 2.9: Correlations between the variables from Table 2.5



Notes: The figure depicts the correlation coefficients between variables included in the benchmark BMA exercise from Table 2.5. The definition and summary statistics of the variables can be found in Table 2.4.

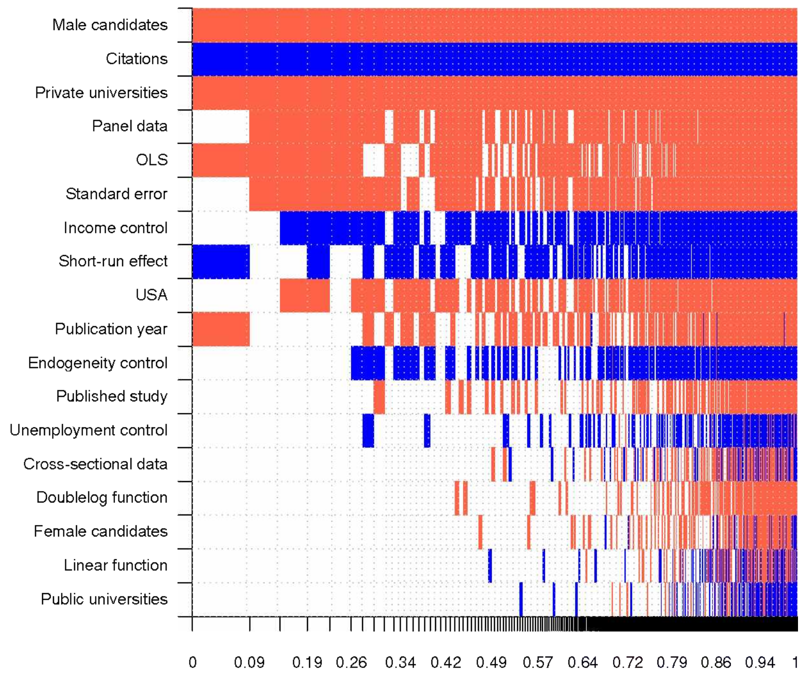
2.B Diagnostics of BMA robustness checks

Table 2.11: Summary of BMA estimation—*Different BMA priors* specification

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.7698	$2 \cdot 10^6$	$1 \cdot 10^5$	5.869949 mins	572,046
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
262,144	21.80%	100%	0.9999	442
Model prior	*g-prior*	*Shrinkage-stats*		
Random	BRIC	Av = 0.9977		

Notes: We employ the "random" model prior, which refers to the beta-binomial prior advocated by Ley & Steel (2009); Zellner's g prior is set according to Fernandez *et al.* (2001). The results of this BMA exercise are reported in Table 2.7 (*different BMA priors* specification).

Figure 2.10: Model inclusion in BMA—*Different BMA priors* specification



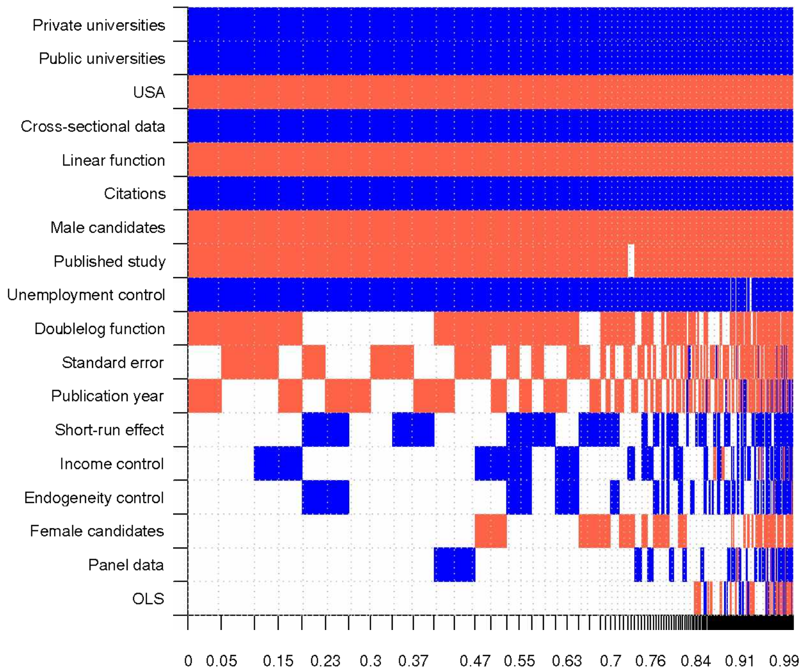
Notes: The figure depicts the results of the BMA related to *different BMA priors* specification reported in Table 2.7.

Table 2.12: Summary of BMA estimation—*Precision-weighted data* specification

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
12.0906	$2 \cdot 10^6$	$1 \cdot 10^5$	6.009023 mins	465,235
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
262,144	17.70%	100%	0.9995	442
Model prior	*g-prior*	*Shrinkage-stats*		
Uniform	UIP	Av = 0.9977		

Notes: We employ the priors suggested by Eicher *et al.* (2011), who recommend using the uniform model prior (each model has the same prior probability) and the unit information prior (the prior provides the same amount of information as one observation in the data). The results of this BMA exercise are reported in Table 2.7 (*precision-weighted data* specification).

Figure 2.11: Model inclusion in BMA—*Precision-weighted data* specification



Notes: The figure depicts the results of the BMA related to *precision-weighted data* specification reported in Table 2.7.

Chapter 3

Publication and Attenuation Biases in Measuring Skill Substitution

Abstract

A key parameter in the analysis of wage inequality is the elasticity of substitution between skilled and unskilled labor. We show that the empirical literature is consistent with both publication and attenuation bias in the estimated inverse elasticities. Publication bias, which exaggerates the mean reported inverse elasticity, dominates and results in corrected inverse elasticities closer to zero than the typically published estimates. The implied mean elasticity is 4, with a lower bound of 2. Elasticities are smaller for developing countries. To derive these results, we use nonlinear tests for publication bias and model averaging techniques that account for model uncertainty.

Keywords: Elasticity of substitution, skill premium, meta-analysis, model uncertainty, publication bias

JEL Codes: J23, J24, J31

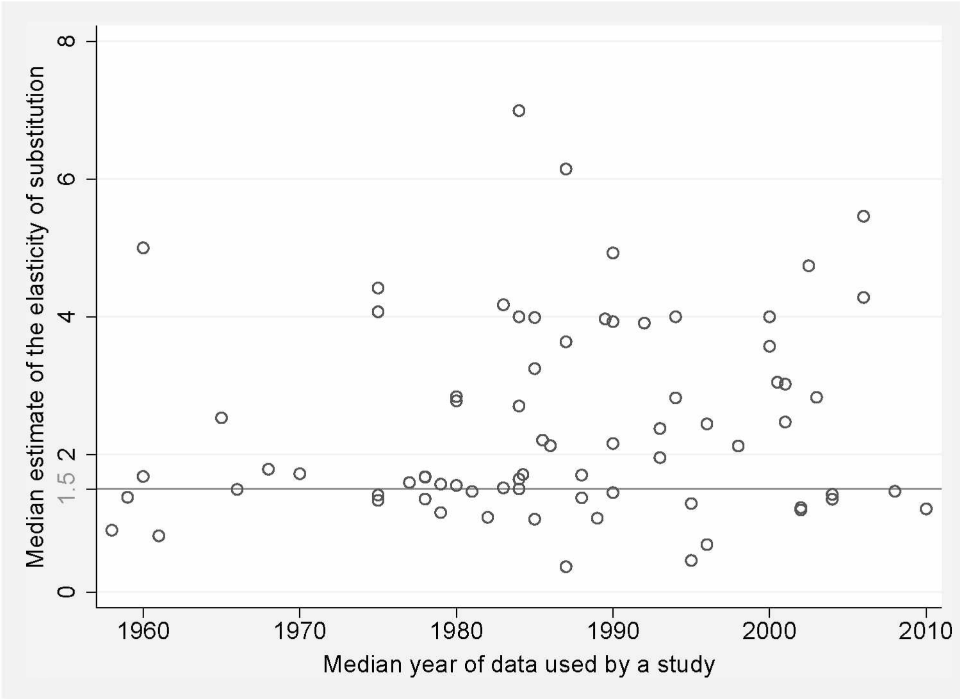
This paper is a joint work with Tomas Havranek, Zuzana Irsova and Lubica Laslopova. The paper is published in the Review of Economics and Statistics. We thank the Editor (Xiaoxia Shi) and Pedro Bom, Chris Doucouliagos, Dominika Ehrenbergerova, Chishio Furukawa, Sebastian Kranz, Heiko Rachinger, Bob Reed, Tom Stanley, Jan Visek, and two anonymous referees for their comments, which helped us improve the paper dramatically.We acknowledges support from the NPO Systemic Risk Institute (grant #LX22NPO5101); from the Czech Science Foundation (grant #19-26812X) and (grant #21-09231S). Data and code are available at `meta-analysis.cz/skill`.

3.1 Introduction

The elasticity of substitution between skilled and unskilled workers ranks among the most frequently estimated parameters in labor economics: we found 682 estimates reported in 77 studies. The parameter commands the predictions of the canonical model of skill differentials, especially the effect on the skill premium of a changing ratio of skilled workers and biased technological change (for instance, Katz & Murphy 1992; Acemoglu 2002; Ciccone & Peri 2005). It is also important for other questions, including the usefulness of cross-country heterogeneity in education for explaining differences in labor productivity (Klenow & Rodriguez-Clare 1997). Unlike many important parameters in economics, for which often little consensus exists and calibrations vary by the order of magnitude, the elasticity of skill substitution is with extraordinary consistency commonly calibrated at 1.5. As Cantore *et al.* (2017, p. 80) put it: "Most of [the] estimates [of the elasticity] range between 1.3 and 2.5, with a consensus estimate around 1.5." In this paper we show that the literature is instead consistent with an elasticity around 4.

The observation by Cantore *et al.* (2017) is based on key papers (Katz & Murphy 1992; Ciccone & Peri 2005; Autor *et al.* 2008) but, at first glance, holds for the literature as a whole: the 682 estimates we collect have a mean of 1.8. Nevertheless, Figure 3.1 illustrates that individual studies estimating the elasticity disagree more than what is often acknowledged in the applications of the estimates. Elasticities larger than 1 (suggesting that skilled and unskilled labor are gross substitutes) dominate the literature and also frequently include values around 4. Elasticities smaller than 1 (suggesting that skilled and unskilled labor are gross complements) are not rare. So the literature is consistent with a wide range of calibrations, though of course the first moment is key in informing them. The problem is that the mean estimates reported in many fields of economics are routinely distorted by publication bias (Brodeur *et al.* 2016; Bruns & Ioannidis 2016; Card *et al.* 2018; Christensen & Miguel 2018; DellaVigna *et al.* 2019; Blanco-Perez & Brodeur 2020; Brodeur *et al.* 2020; Ugur *et al.* 2020; Xue *et al.* 2020; Imai *et al.* 2021; Neisser 2021; Stanley *et al.*

Figure 3.1: Many studies defy the consensus of 1.5 elasticity



Notes: The vertical axis shows the median estimate of the elasticity of substitution reported in individual studies. The horizontal axis shows the median year of the data used in the studies. Outliers are omitted from the figure for ease of exposition but included in all tests. The figure, as well as all other figures, tables, and numbers in the main text, only considers elasticities implied by regressions of the skill premium on the relative supply of skilled labor, not elasticities implied by reverse regressions (see text and Appendix 3B for details.

2021; Brown *et al.* 2022; DellaVigna & Linos 2022; Iwasaki 2022; Stanley *et al.* 2022), often by a factor of 2 or more (Ioannidis *et al.* 2017).

Publication bias stems from the tendency of authors, editors, or referees to prefer statistically significant or theory-consistent results. Negative estimates of the elasticity are inconsistent with the canonical model, and zero or infinite estimates are unintuitive. Few researchers are eager to interpret such estimates, though negative, insignificant, or huge elasticity estimates will appear from time to time given sufficient imprecision in data and methods. The analysis of publication bias in this context is complicated by the fact that while some researchers estimate the elasticity directly, most estimate the (negative) inverse elasticity by regressing the skill premium on the relative supply of skilled labor. The two groups of studies cannot be combined in an analysis of publication bias because the inversion necessary for such a combination violates the assumptions of many tests. Since in most plausible situations the relative supply represents the treatment and the skill premium represents the outcome, in the main text we only focus on the studies estimating the negative inverse elasticity, which are more likely to identify the underlying causal relationship. In the Appendix we explain in detail why we find direct estimates, yielded by reverse regressions, less persuasive (Appendix 3B), and provide tests of publication bias for these estimates separately (Appendix 3C). The direct estimates are consistent with little to no substitutability between skilled and unskilled labor.

McCloskey & Ziliak (2019) liken the problem of publication bias and p-hacking[1] to the Lombard effect in psychoacoustics, in which speakers intesify their vocal effort in response to noise. So, too, can researchers intensify specification searching in response to noise in their data and try a different setup to obtain a negative inverse elasticity larger in magnitude, ideally an estimate significantly different from zero. Most of the techniques we use for publication bias correction (including Ioannidis *et al.* 2017; Andrews & Kasy 2019; Bom & Rachinger 2019; Furukawa 2020) are explicitly or implicitly based on the Lombard effect and assume that, in the absence of the bias, there is no correlation between estimates and standard errors. The assumption is common but strong, and we show that the correlation exists even among estimates unlikely to suffer from the bias. Consequently we use the inverse of the square root of the number of observations as an instrument for the standard error (Stanley 2005) and employ tests by Gerber & Malhotra (2008) and Elliott *et al.* (2022) that do not require the assumption.

We have noted that publication bias has been identified in many fields. In most cases, however, it is probably moderated by attenuation bias in the opposite direction. According to the "iron law of econometrics" (Hausman 2001), most estimates are biased towards zero because the independent variable is almost always measured with error. The interplay between publication and attenuation biases must be ubiquitous

[1] Conceptually, publication bias and p-hacking are distinct terms. The latter denotes researchers' effort to produce statistically significant results, and often stems from publication bias. But it is unfeasible in empirical work to separate these two effects, as they tend to be observationally equivalent. Applied meta-analysts thus typically use the term publication bias more generally to also include p-hacking, and we follow this practice.

in economics, but to our knowledge has not been explored before. The literature on skill substitution recognizes the measurement error problem, since data on labor supply can be notoriously noisy, and attenuation bias is mentioned frequently (e.g. by Katz & Murphy 1992; Angrist 1995; Borjas 2003; Bound *et al.* 2004; Borjas & Katz 2007; Autor *et al.* 2008; Card 2009; Behar 2010; Verdugo 2014; Kawaguchi & Mori 2016; Bowlus *et al.* 2022). A classical measurement error can arise in the relative labor supply for at least three reasons. First, survey responses may contain noise. Second, migrants' degrees may be incomparable to natives' degrees due to cross-country differences in the quality of the educational system. Third, the mapping from degrees to skills may be noisy due to time differences in the quality of education and selection into student cohorts. We exploit the fact that part of the literature uses instrumental variables (IV) to address the attenuation bias and other endogeneity biases, while other studies either use simple OLS or have access to arguably exogenous variation in relative labor supply (natural experiments). The differences in results reported for studies based on OLS, IV, and natural experiments are informative on the extent of attenuation bias.

Our results are consistent with both publication and attenuation bias. After correcting for the former, the estimated negative inverse elasticity declines in magnitude from the reported mean of -0.6 to an interval between -0.3 and 0.1, depending on the publication bias correction method. Concerning the latter, the publication bias corrected mean estimates are close to zero for both OLS and natural experiments, but around -0.25 for IV. Under the assumption that the instrumental variables in the literature are generally specified well, this result suggests that attenuation bias or other endogeneity biases are important on average (the difference between OLS and IV is substantial) and that attenuation bias in particular matters (the difference between IV and natural experiments is substantial, too). Our preferred estimate of the mean elasticity is thus 4, a value approximately corrected for both publication and attenuation bias.

The results are corroborated by a model that controls for 24 characteristics that reflect the context in which the estimates were obtained (for example, variable definition, data characteristics, design of the production function, estimation technique, and publication characteristics). To address the resulting model uncertainty we use Bayesian (Raftery *et al.* 1997; Eicher *et al.* 2011) and frequentist (Hansen 2007; Amini & Parmeter 2012) model averaging, both superbly surveyed in Steel (2020). For the former we also employ the dilution prior (George 2010) that alleviates potential collinearity. Finally, we create a hypothetical study that uses all estimates in the literature but assigns more weight to those that are better specified (using Card, 2009, Autor, 2014, and Carneiro *et al.*, 2022, as benchmarks). The implied mean estimate of the elasticity is 4 with the 95% credible interval of $(2, 20)$. The implied elasticity for the US is 6, and for developing countries it is 2. We also find that publication bias is smaller for IV estimates and developing countries, likely because for them the underlying inverse elasticity estimates are significantly distinct from zero even in the absence of publication selection.

The remainder of the paper contains an analysis of publication bias (Section 3.2) and heterogeneity (Section 3.3); attenuation bias is analyzed in both sections. The Appendix provides details on the dataset and estimation of the elasticity (Appendix 3A), discussion of the studies estimating the elasticity directly (Appendix 3B), additional material on publication bias analysis (Appendix 3C), additional material on heterogeneity analysis (Appendix 3D), and diagnostics and robustness checks of the Bayesian model averaging analysis (Appendix 3E). Data and code are available at `meta-analysis.cz/skill`.

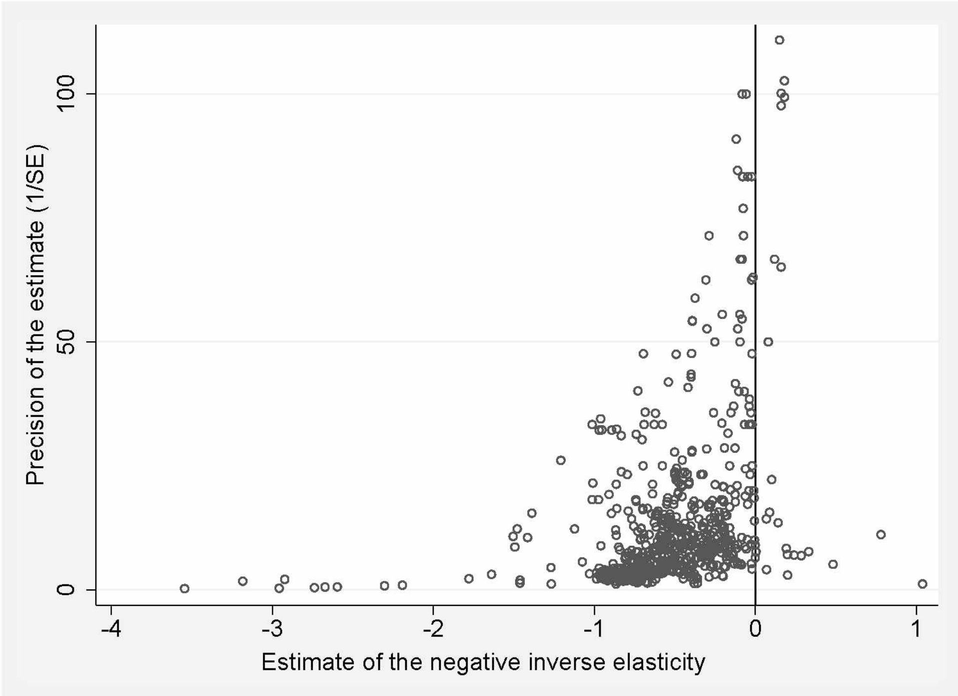
3.2 Publication Bias

An intuitive quality of the elasticity of substitution between skilled and unskilled labor is its nonnegativity. As Kearney (1997, p. 33) remarks on his negative estimates: "The implied coefficients ... violate standard economic theory." Some researchers, such as Bowles (1970, p. 73) "exclude [negative estimated] values [of the elasticity] ... as implausible on a priori grounds." As we have noted, we focus on studies that estimate the (negative) inverse elasticity. An inverse elasticity of zero, implying infinite elasticity of substitution, is theoretically possible but often deemed implausible and rarely interpreted. What follows is a tendency in the literature to discriminate against positive and insignificant values of the negative inverse elasticity. Hence the mean estimate of the negative inverse elasticity is probably biased towards a negative value larger in magnitude. Such publication bias is natural, inevitable, and does not require any ulterior motives on the side of authors, editors, or referees. It is a task for those who review and interpret the literature to correct for the bias. As far as we know, no one has attempted to do so in the case of the elasticity of skill substitution.

Most tests of publication bias assume that in the absence of the bias there is no correlation between reported estimates and their standard errors. The correlation can capture publication bias for two reasons. First, researchers (or editors or referees) may prefer statistically significant results. Given some imprecision in their data and methods, researchers may try, for example, different combinations of control variables until they obtain an estimate large enough to offset the standard error. Second, researchers may prefer an intuitive sign of the estimates and discard those with the opposite sign. Then correlation between estimates and standard errors arises due to heteroskedasticity: with lower precision, estimates will be more dispersed on both sides of the underlying mean elasticity. When positive estimates of the negative inverse elasticity are discarded, a regression of estimates on standard errors will yield a negative slope coefficient.

It is helpful to evaluate the relationship visually using the so-called funnel plot: a scatter plot of estimates on the horizontal and their precision (1/SE) on the vertical axis. Based on the intuition described in the previous paragraph, an asymmetry of the funnel plot suggests publication bias, and the top of the funnel serves as an indication of the underlying mean elasticity corrected for the bias. This is the case because under the assumption that all studies estimate the same underlying elasticity the most precise estimates are likely to be close to the underlying mean; moreover, because of their high precision they tend to be highly significant and less prone to publication bias. Figure 3.2 shows evidence consistent with implicit or explicit discrimination against estimates with the unintuitive (positive) sign. The

Figure 3.2: The funnel plot suggests publication bias



Notes: In the absence of publication bias the funnel plot should be symmetrical. Outliers are excluded from the figure for ease of exposition but included in all statistical tests. SE = standard error.

most precise estimates are concentrated around zero, which is consistent with perfect substitutability between skilled and unskilled labor.

We use two groups of tests more formal than the funnel plot. First, we regress estimates on their standard errors and, to address heteroskedasticity, weight the regressions by inverse variance in the spirit of Stanley (2008), Doucouliagos & Stanley (2013), and Stanley & Doucouliagos (2015). Second, we use recent techniques that do not rely on the linearity assumption. Regarding the linear meta-regression, a nonzero estimated slope suggests publication bias. Under the assumption that publication selection is a linear function of the standard error and there is no heterogeneity in the literature, the intercept can be interpreted as the true mean elasticity corrected for the bias (the top of the funnel). The linearity assumption, however, cannot be expected to hold in general, as explained by Andrews & Kasy (2019) in the appendix to their paper (pp. 30–31).

Regarding nonlinear models, the technique with the most rigorous foundations is the selection model of Andrews & Kasy (2019), which estimates the probability of a result being reported and uses the probability to re-weight the observed distribution of results. We have to specify the thresholds for the t-statistic associated with changes in publication probability, and we choose -1.96, 0, and 1.96.[2] We assume that effects have a t-distribution and we cluster standard errors at the study level. The other nonlinear specification that we employ is the endogenous kink model by Bom & Rachinger (2019), which builds on Stanley & Doucouliagos (2014). It assumes that the relation between estimates and standard errors is linear up to a certain point until when precision is high enough for all estimates to be published and the relation disappears. The endogenous kink technique represents the latest incarnation of tests based directly on the funnel plot.

While the nonlinear techniques do not use the problematic assumption that publication selection is a linear function of the standard error, they share the strong assumption that estimates and standard errors are independent or at least uncorrelated in the absence of bias. Andrews & Kasy (2019) state the independence assumption explicitly, while the endogenous kink technique implicitly assumes that more precise estimates are less biased and closer to the true value.[3] The assumption is unlikely to hold in economics because data and method choices can influence both estimates and standard errors systematically. Table 3.15 in the Appendix shows that estimates and their standard errors are correlated even among estimates with a p-value below 0.005, where publication bias is less likely. The correlation appears in most cases even if we divide the literature to subsamples according to the main differences in data and methods. But it is also possible that even these highly significant estimates are plagued by publication bias.

Table 3.16 in the Appendix presents a direct specification test, introduced by Kranz & Putz (2022) on the suggestion of Isaiah Andrews, of the Andrews & Kasy

[2] We only report the probability related to the -1.96 threshold for negative inverse estimates; some of the remaining groups (especially positive estimates of the negative inverse elasticity) have a limited number of observations.

[3] If there is, for example, a positive relationship between estimates and standard errors in the absence of publication bias, highly precise estimates will be smaller than the true underlying mean. If some researchers reduce standard errors (for example, via changes in clustering) in response to small point estimates, high reported precision can be spurious.

(2019) technique. The table shows, for various subsets of the literature, 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 & Kasy (2019) model. If all the assumptions of the model hold, the correlation should be zero. In our case the correlation is substantial for almost all subsets of the literature, which means that some of the assumptions (including the key independence assumption) are probably violated.

As a partial solution to the likely violation of the independence assumption invoked by nearly all meta-analysis techniques, we run a simple meta-regression where the standard error is instrumented by the inverse of the square root of the number of observations (Stanley 2005; Havranek 2015). Comparing this IV estimate with other linear and nonlinear estimators tells us something about the practical importance of the independence assumption for measuring the magnitude of publication bias and the corrected effect. Following Andrews *et al.* (2019), we report the two-step weak-instrument-robust 95% confidence interval based on the Stata package by Sun (2018) and the idea of Andrews (2016) and Andrews (2018).

In the main text we focus on 5 bias-correction estimators that we consider most informative in the context of skill substitution: linear meta-regression with study-level fixed effects, between-effects meta-regression, IV meta-regression, the Bom & Rachinger (2019) endogenous kink model, and the Andrews & Kasy (2019) selection model. In the Appendix we also report the results of three additional techniques: OLS meta-regression, the weighted average of adequately powered estimates introduced by Ioannidis *et al.* (2017), and the stem-based technique by Furukawa (2020). The results of these three techniques generally do not alter our conclusions. Each of the 5 estimators that we focus on has a different strength: the fixed-effects model allows us to filter out idiosyncratic study-level effects, the between-effects model gives each study the same weight, the IV meta-regression directly addresses potential endogeneity, the endogenous kink model is the most advanced nonlinear estimator based on the funnel plot and performs well in Monte Carlo simulations (Bom &

Rachinger 2019), and the Andrews & Kasy (2019) model is the one most rigorously founded, although, as we have noted, in the case of skill substitution probably not well specified.

In the Appendix (Table 3.10) we test publication bias for the entire sample of negative inverse elasticity estimates. All techniques find substantial publication bias and, with the exception of the Andrews & Kasy (2019) model, yield estimated mean inverse elasticities close to zero.[4] Even for the Andrews & Kasy (2019) model the implied mean elasticity of substitution exceeds 3. In the main text we analyze publication bias separately for different methods used in the primary studies and divide the studies into three groups: OLS (typically time series studies that either ignore endogeneity or argue that it is not a major issue), IV (typically cross-sectional studies with shift-share instruments), and natural experiments (studies that exploit arguably exogenous variation in relative skill supply induced either by migration or expansions of higher education).

Correcting for publication bias in individual subsamples separately has three advantages. First, the aggregate analysis may confound publication bias with heterogeneity. Second, previous meta-analyses have shown differences in publication bias between OLS and IV estimates in economics. For example, Ashenfelter *et al.* (1999) find that IV estimates of the return to schooling suffer more from publication bias because researchers have a harder time producing statistically significant estimates given the imprecision brought by IV. Third, differences in the corrected means for OLS, IV, and natural experiments are informative on the extent of attenuation bias. If IV studies are well specified, they correct for attenuation bias and other endogeneity biases. Natural experiments correct for other endogeneity biases, but in general not for attenuation bias.

Table 3.1 shows the results. For natural experiments we only have 40 estimates

[4]For the sample of direct elasticity estimates we also find strong publication bias and zero mean corrected coefficient. Thus both groups of studies suggest little correlation between the wage premium and relative labor supply. But inference regarding the elasticity is the opposite for the two groups. As explained in the Appendix (Appendix 3B), we find less persuasive the identification arguments used by studies estimating the elasticity directly. Moreover, there are not enough IV and natural experiment studies on direct estimates to allow us examine attenuation bias for direct estimates.

taken from 6 studies, so the power of the tests is low for this group, but all techniques suggest strong publication bias and negligible corrected effects. Natural experiments as a whole are thus consistent with no causal effect of relative skill supply on the skill premium and therefore with infinite elasticity of substitution. We obtain similar results for OLS estimates—with the exception of the Andrews & Kasy (2019) model, which is in this context less aggressive in correcting for publication bias. But IV estimates of the negative inverse elasticity are different: they show less publication bias and larger corrected inverse elasticities, implying the elasticity of substitution around 4. The results are consistent with attenuation bias in the literature (IV estimates of negative inverse elasticities are larger in magnitude than OLS estimates) and little additional endogeneity bias (OLS estimates are similar to estimates from natural experiments). Nevertheless, even our preferred estimate of 4 is much larger than the uncorrected mean implied elasticity of 1.8, a difference which shows that publication bias dominates attenuation bias. In contrast to Ashenfelter *et al.* (1999), we find that IV estimates suffer less from publication bias than OLS estimates.[5] This is the case because the underlying inverse elasticity is much farther from zero for IV relative to OLS estimates, which means that with IV less effort is needed to obtain plausible estimates for publication.

In the Appendix (Table 3.11, Table 3.12, Table 3.13) we test and correct for publication bias in other variously defined subsamples of the literature: elasticities estimated for developed countries vs. elasticities for developing countries, elasticities estimated at the country level vs. elasticities at the regional level, and elasticities estimated using a one-level CES function vs. a multilevel CES function. The results suggest that elasticities tend to be larger for developed countries (above 4) than developing countries (around 2.5), and once again publication bias is stronger for the group which displays a corrected inverse elasticity closer to zero.

[5] Our findings also contrast those of Brodeur *et al.* (2020), who find that IV estimates are more biased than other techniques commonly used in economics. But note that Brodeur *et al.* (2020) only examine (quasi-)experimental techniques (IV, difference-in-differences, regression discontinuity design, randomized control trials), not OLS.

Table 3.1: IV estimation of the negative inverse elasticity shows less bias and a larger corrected effect in magnitude compared to both OLS and natural experiments

Panel A: OLS estimates	FE	BE	IV	EK	SM
Publication bias	-5.804***	-4.277***	-6.962***	-5.465***	P=0.468
	(1.999)	(1.266)	(1.694)	(0.540)	(0.139)
			[-11.770, -2.494]		
			{-11.972, -3.133}		
Effect beyond bias	-0.021	-0.097	0.010	-0.036**	-0.289**
	(0.103)	(0.063)	(0.104)	(0.019)	(0.113)
			[-0.331, 0.214]		
First-stage robust F-stat			46.17		
Observations	347	347	251	347	347
Panel B: IV estimates	FE	BE	IV	EK	SM
Publication bias	-2.287**	-0.923	-0.553	-1.485***	P=0.336
	(0.843)	(1.365)	(0.681)	(0.268)	(0.093)
			[-1.913, 1.078]		
			{-1.991, 0.748}		
Effect beyond bias	-0.149	-0.297**	-0.400***	-0.252***	-0.333***
	(0.109)	(0.115)	(0.114)	(0.025)	(0.058)
			[-0.719, 0.175]		
First-stage robust F-stat			69.98		
Observations	264	264	212	264	264
Panel C: Natural experiment estimates	FE	BE	IV	EK	SM
Publication bias	-3.557***	-1.874*	-3.176***	-3.115***	P=0.187
	(0.018)	(0.682)	(0.853)	(0.343)	(0.075)
			[-4.854, -1.407]		
			{-4.653, -1.444}		
Effect beyond bias	0.0496***	-0.121	-0.003	0.003	-0.009
	(0.003)	(0.082)	(0.029)	(0.028)	(0.066)
			[NA, NA]		
First-stage robust F-stat			260.41		
Observations	40	40	40	40	40

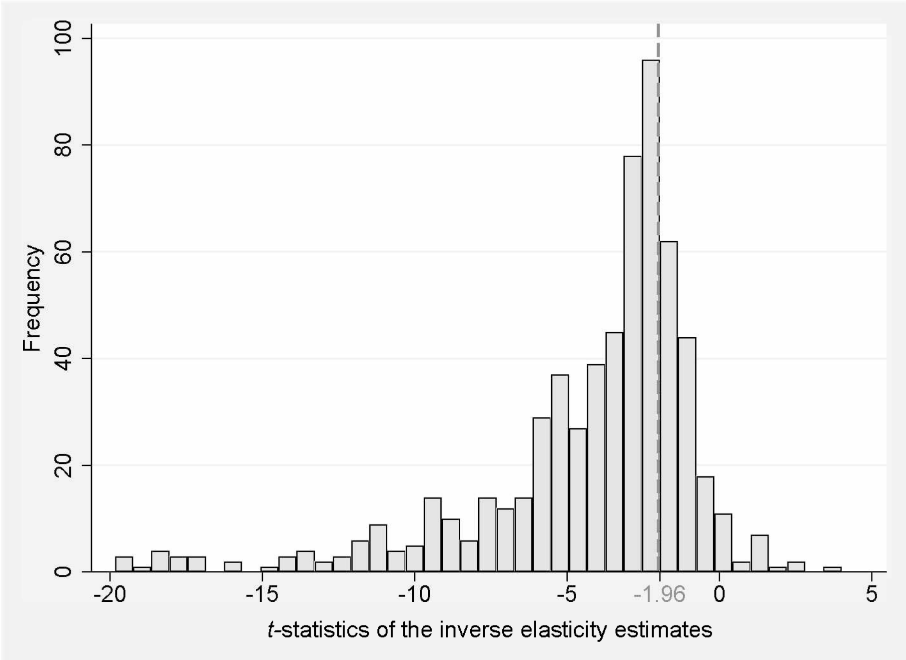
Notes: The first three specifications regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In square brackets we show the 95% confidence interval from wild bootstrap (Roodman *et al.* 2018); in curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). EK = endogenous kink method by Bom & Rachinger (2019), SM = selection model by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The cross-country differences in elasticities are discussed, for example, by Behar (2010). A plausible explanation for the finding is that in many developing countries access to higher education is still limited, and therefore selection effects are stronger within cohorts. In addition, the unskilled labor aggregate contains workers of limited literacy. Next, our results suggest that elasticities estimated at the country level are smaller than those estimated at the regional level, but there are only 93 estimates for the latter group. Finally, both one-level and multilevel CES functions seem to yield similar estimated elasticities.

In addition to bias-correction methods, we use the caliper test for the distribution of t-statistics by Gerber & Malhotra (2008) and two new tests for the distribution of p-values developed by Elliott *et al.* (2022). These tests of publication bias do not need the independence assumption, but are not designed to estimate the underlying elasticity. Figure 3.3 provides a motivation: the frequency of reported estimates drops precipitously when the t-statistic falls short of -1.96 in magnitude. The first block of Table 3.2 examines this drop using the caliper test (Gerber & Malhotra 2008). In a narrow caliper around -1.96, 62% of the estimates are different from zero at the 5% level, while only 38% of them are statistically insignificant. In the histogram of the estimates (Figure 3.5 in the Appendix) we observe that, in addition to 0, -1 is an important threshold. It is unintuitive to suggest that skilled and unskilled labor are gross complements, and the value -1 itself would mean that skill-biased technical change has no effect on the skill premium. In the second block of the table we thus test whether authors prefer to report estimates rejecting a negative inverse elasticity of -1. In this case the caliper test is inconclusive. Next, we look at the distribution of inverted elasticities itself, not t-statistics, and confirm the large drops at 0 and -1 as apparent from Figure 3.5.

The disadvantage of caliper tests is the necessity to specify the values where we expect breaks in the distribution. Elliott *et al.* (2022) derive two new rigorously founded techniques that do not require us to define the location of the breaks. The techniques rely on the conditional chi-squared test of Cox & Shi (2022). The first technique is a histogram-based test for non-increasingness of the p-curve, the second

Figure 3.3: The distribution of t-statistics peaks at -2



Notes: The dashed vertical line represents the critical value associated with significance at the 5% level. For ease of exposition we exclude outliers from the figure but include them in all statistical tests.

technique is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives. In their applications, Elliott *et al.* (2022) only focus on p-values below 0.15 and use 15, 30, or 60 bins. Because our dataset is much smaller (especially in subsamples), we include all p-values below 0.2 and use 5–10 bins depending on the size of the subsample. In most cases we reject the null hypothesis of no publication bias, with the exception of natural experiments, regional estimates, and developing countries. These are also the smallest subsamples, which might suggest that larger datasets than ours are needed for the tests of Elliott *et al.* (2022) to have adequate power.

Table 3.2: Tests based on the distribution of t-statistics and p-values

Panel A: Caliper tests due to Gerber & Malhotra (2008)				
Threshold for t*-statistic:* -1.96	caliper: 0.25	0.30	0.35	0.40
Share above threshold minus 0.5	-0.118**	-0.135**	-0.102**	-0.121***
	(0.0561)	(0.0525)	(0.0485)	(0.0452)
Observations	76	85	103	116
Threshold for adjusted t*-statistic* $t^* = (estimate + 1)/SE(estimate)$*:* 1.96 *(relevant for the null hypothesis that the negative inverse elasticity is* -1*)*				
	caliper: 0.25	0.30	0.35	0.40
Share above threshold minus 0.5	0.090	0.100	0.088	0.096
	(0.0798)	(0.0739)	(0.0696)	(0.0656)
Observations	39	45	51	57
Threshold for neg. inv. elasticity: 0	caliper: 0.05	0.10	0.15	0.20
Share above threshold minus 0.5	-0.397***	-0.387***	-0.379***	-0.383***
	(0.0492)	(0.0439)	(0.0405)	(0.0369)
Observations	39	53	66	77
Threshold for neg. inv. elasticity: -1	caliper: 0.05	0.10	0.15	0.20
Share above threshold minus 0.5	0.346***	0.368***	0.378***	0.406***
	(0.0722)	(0.0556)	(0.0473)	(0.0367)
Observations	26	38	49	64

Panel B: Tests due to Elliott *et al.* (2022)					
	All inverse	OLS method	IV method	Natural experiment	Developed country
Test for non-increasingness	0.016	0.037	0.307	1.000	0.098
Test for monotonicity and bounds	0.008	0.050	0.032	1.000	0.110
Observations ($p <= 0.2$)	586	315	230	39	369
Total observations	654	347	264	40	418
	Developing country	Country estimate	Region estimate	One-level CES	Multilevel CES
Test for non-increasingness	1.000	0.078	1.000	0.000	0.025
Test for monotonicity and bounds	0.930	0.041	0.773	0.000	0.016
Observations ($p <= 0.2$)	138	491	89	173	403
Total observations	151	555	93	198	444

Notes: In Panel A, the tests compare the relative frequency of estimates above and below an important threshold for the t-statistic or negative inverse elasticity. A test statistic of -0.397, for example, means that 89.7% estimates are below the threshold and 10.3% estimates are above the threshold. Panel B reports for different subsamples the p-values of two tests developed by Elliott *et al.* (2022), which also feature cluster-robust variance estimators. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.3 Heterogeneity

The literature on the elasticity of substitution is characterized by significant variation in the reported estimates, as we have shown in Figure 3.1. While publication bias explains a part of this variation, individual studies (and individual specifications within the studies) differ greatly in terms of the data and methods used. In this section we control for 24 variables that capture the context in which researchers obtain their estimates. Given the model uncertainty inherent in such an exercise, we use Bayesian and frequentist model averaging. Our goals are threefold. First, we examine whether the relation between estimates and standard errors, which serves as an indication of publication bias, is robust to controlling for the aspects of study design. This analysis complements the IV meta-regression approach presented in the previous section. Second, we aim to identify the aspects that are the most effective in explaining the differences among the reported elasticities. Third, as the bottom line we create a synthetic study that computes an implied elasticity using all estimates but giving more weight to those that are arguably better identified and correcting for both publication and attenuation bias.

Table 3.3 lists the variables that we use; they are described in more detail, including motivation for their inclusion, in Table 3.17 and Appendix 3D in the Appendix. We divide the variables into five groups: data characteristics (such as data frequency and aggregation), structural variation (different countries and sectors), production function design (for example, one-level vs. multilevel specifications), estimation technique (for example, OLS vs. IV vs. natural experiments), and publication characteristics (impact factor of the outlet and the number of citations received per year). The latter group is included as a proxy for quality not captured by the data and method characteristics. As explained in Appendix 3D, some of the dummy variables are used as reference categories, so they are not all included in regressions.

In addition, we include interactions of the standard error and the dummy variables for IV estimates and developing countries, respectively, because the results in the previous section suggest that the corresponding estimates are less affected by

Table 3.3: Characteristics used to explain heterogeneity

Category	Variables
Data characteristics:	Annual frequency, Higher frequency, Lower frequency, Micro data, Sectoral data, Aggregated data, Cross-section
Structural variation:	United States, Developing country, Manufacturing sector
Design of the production function:	One-level CES function, Multilevel CES function, Time control, Location control, Macro control, Age control, Capital control
Estimation technique:	Dynamic model, Unit fixed effects, Time fixed effects, OLS method, IV method, Natural experiment
Publication characteristics:	Impact factor, Citations

Notes: Details on each variable, including definition, summary statistics, and motivation for inclusion, are available in Table 3.17 and Appendix 3D in the Appendix. In data collection we follow the guidelines compiled by the Meta-Analysis in Economics Research Network (Havranek *et al.* 2020).

publication bias. That leaves 24 variables in total for all models in this section.

Ideally we would regress the collected inverse elasticities on the 24 variables described above. Given such a large number of regressors, however, the probability that many will prove redundant is high, which would compromise the precision of parameter estimates for the more important regression variables. In other words, we face substantial model uncertainty; to address it, we employ model averaging techniques, both Bayesian and frequentist. The Bayesian approach allows us to estimate the probability that an individual explanatory variable should be included in the underlying model. The frequentist approach is computationally more cumbersome, but does not require the choice of priors and serves as a useful robustness check.

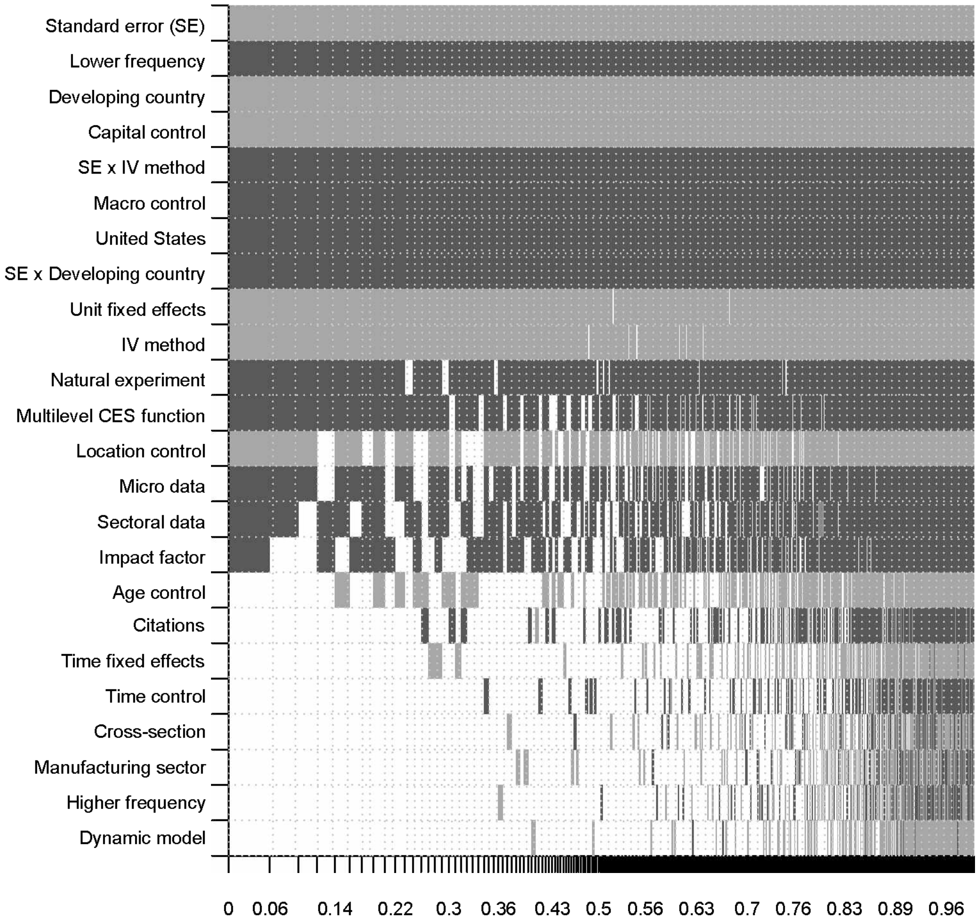
The goal of Bayesian model averaging (BMA) is to find the best possible approximation of the distribution of regression parameters. The method yields three basic statistics for each parameter: posterior mean, posterior variance, and posterior inclusion probability. In our case BMA is to run 2^{24} regressions determined by all the possible combinations of the explanatory variables. We simplify this task by employing the Metropolis-Hastings algorithm of the `bms` package for R by Zeugner & Feldkircher (2015), which walks only through the most likely models. The likelihood of each model is reflected by posterior model probabilities (analogous to information criteria in the frequentist setting). Posterior means are then computed as the estimated coefficients weighted across all models by their posterior model probability. The posterior inclusion probability of a variable is defined as the sum of posterior model probabilities for all models where this candidate regressor is included (analogous to statistical significance in the frequentist setting). For more details on BMA, we refer the reader to Raftery *et al.* (1997) and Eicher *et al.* (2011); BMA has already been used in meta-analysis by Bajzik *et al.* (2020), Zigraiova *et al.* (2021), Gechert *et al.* (2022), and Matousek *et al.* (2022).

BMA requires explicit priors concerning the model (model prior) and regression coefficients (g-prior). Our baseline model prior and g-prior reflect our lack of ex ante information in both areas: we employ a uniform model prior, which gives each model the same prior probability, and the unit information g-prior, which provides the same information as one observation from the data (suggested by Eicher *et al.* 2011). In addition, we employ the dilution prior according to 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.

Furthermore, in the Appendix (Appendix 3E) we combine the random model prior (following Ley & Steel 2009) with the hyper-g prior (suggested by Feldkircher & Zeugner 2012): while the random model prior assumes that the distribution of the model size to be beta-binomial (which reflects the fact that no model *size* is preferred), the hyper-g prior sets the prior expected shrinkage factor equivalent to the BRIC parameter prior (see Fernandez *et al.* 2001, suggesting multivariate normal distribution that has a covariance matrix specified depending on the data). In our application of frequentist model averaging we use Mallow's weights (Hansen 2007) with orthogonalization of the covariate space according to Amini & Parmeter (2012) to narrow down the number of estimated models. Variables enter the model in descending order by the absolute value of the correlation coefficient with the estimated inverse elasticity. For more details and applications of model averaging techniques in economics, we refer the reader to the superb survey by Steel (2020).

The results of Bayesian model averaging are visualized in Figure 3.4. Each column represents an individual regression model, and the width of the column indicates the corresponding posterior model probability: the weight of the model. The columns are ordered by posterior model probability from left to right in descending order.

Figure 3.4: Model inclusion in Bayesian model averaging



Notes: The variables are sorted according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis measures cumulative posterior model probability. Darker shade of gray color = the estimated parameter for the variable is positive. Lighter shade of gray color = the estimated parameter for the variable is negative. No color = the variable is not included in the model. Numerical results are reported in Table 3.4. All variables are described in Table 3.17 in the Appendix.

Each row of the figure represents a regression variable. The rows are ordered by the posterior inclusion probability from top to bottom in descending order. Each cell with a darker gray color indicates a positive sign of the posterior mean of the regression coefficient for the variable in a given model. Each cell with a lighter gray color indicates a negative sign. If a variable is excluded from the model, the corresponding cell is blank. The figure suggests that approximately two thirds of our explanatory variables are, at least to some degree, useful in explaining the heterogeneity in the reported estimates of the inverse elasticity of substitution; moreover, for these variables the coefficient signs are robust across virtually all the models.

The corresponding numerical results are reported in Table 3.4. The first specification represents our baseline BMA exercise. To interpret the posterior inclusion probabilities (PIPs) of the BMA means, researchers typically follow Jeffreys (1961), who denotes evidence of an effect as 'weak' for a PIP between 0.5 and 0.75, 'substantial' for a PIP between 0.75 and 0.95, 'strong' for a PIP between 0.95 and 0.99, and 'decisive' for a PIP larger than 0.99. The other two specifications in Table 3.4 represent robustness checks: first, ordinary least squares that exclude all the variables deemed utterly unimportant by BMA (with PIP below 0.5); second, frequentist model averaging (FMA) that includes all the variables we have collected. Thus our baseline estimation technique is purely Bayesian, the first robustness check uses Bayesian techniques for the selection of variables but frequentist techniques for estimation, and the second robustness check is purely frequentist. In addition, the Appendix (Appendix 3E) provides more robustness checks that focus on different priors for BMA (Table 3.19).

We focus on the variables for which we have the most robust evidence across the three specifications: at least substantial posterior inclusion probability in Bayesian model averaging and, at the same time, significance at least at the 10% level in both frequentist check and frequentist model averaging. The pre-eminent variable in this respect is the standard error, which shows the strongest association with the reported inverse elasticity in all the models we run.

Table 3.4: Why estimates of the negative inverse elasticity vary

Response variable: Reported estimate	Bayesian model averaging			Frequentist check (OLS)			Frequentist model averaging		
	P.M	P.SD	PIP	Coef.	SE	*p*-val.	Coef.	SE	*p*-val.
Constant	-0.20	NA	1.00	-0.22	0.11	0.04	0.00	0.21	1.00
Standard error (SE)	-3.62	0.84	1.00	-3.60	0.57	0.00	-4.82	1.25	0.00
SE * IV method	2.35	0.48	1.00	2.36	0.74	0.00	2.92	1.18	0.01
SE * Developing country	2.24	0.59	1.00	2.26	0.98	0.02	2.59	1.06	0.01
Data characteristics									
Higher frequency	0.00	0.02	0.08				0.00	0.04	1.00
Lower frequency	0.26	0.04	1.00	0.28	0.09	0.00	0.16	0.11	0.14
Micro data	0.06	0.05	0.65	0.09	0.06	0.15	0.00	0.10	1.00
Sectoral data	0.07	0.06	0.61	0.11	0.08	0.18	0.00	0.11	1.00
Cross-section	0.00	0.01	0.10				0.00	0.03	1.00
Structural variation									
United States	0.10	0.03	1.00	0.10	0.06	0.11	0.02	0.07	0.79
Developing country	-0.21	0.04	1.00	-0.20	0.10	0.05	-0.29	0.14	0.04
Manufacturing sector	0.00	0.02	0.09				0.00	0.03	1.00
Design of production function									
Multilevel CES function	0.05	0.04	0.79	0.07	0.08	0.37	-0.02	0.08	0.83
Time control	0.00	0.01	0.11				0.00	0.00	1.00
Location control	-0.10	0.08	0.65	-0.14	0.10	0.15	0.00	0.14	1.00
Macro control	0.19	0.04	1.00	0.21	0.06	0.00	0.04	0.16	0.81
Age control	-0.02	0.03	0.36				0.00	0.03	1.00
Capital control	-0.39	0.03	1.00	-0.39	0.09	0.00	-0.42	0.13	0.00
Estimation technique									
Dynamic model	0.00	0.02	0.07				0.00	0.01	1.00
Unit fixed effects	-0.08	0.02	0.99	-0.09	0.04	0.02	-0.02	0.06	0.72
Time fixed effects	0.00	0.01	0.13				0.00	0.02	1.00
IV method	-0.12	0.04	0.96	-0.13	0.07	0.06	-0.12	0.05	0.02
Natural experiment	0.19	0.08	0.92	0.18	0.07	0.01	0.13	0.10	0.20
Publication characteristics									
Impact factor	0.01	0.01	0.55	0.02	0.02	0.40	0.00	0.02	1.00
Citations	0.00	0.01	0.20				0.00	0.00	1.00
Studies	68			68			68		
Observations	654			654			654		

Notes: P.M = posterior mean, P.SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. In Bayesian model averaging we employ the combination of the uniform model prior recommended by Eicher *et al.* (2011) and the dilution prior (George 2010), which accounts for collinearity. The frequentist check (OLS) includes the variables found by BMA to have PIP above 0.5 and is estimated using standard errors clustered at the study level. Frequentist model averaging applies Mallow's weights (Hansen 2007) using orthogonalization of covariate space suggested by Amini & Parmeter (2012) to reduce the number of estimated models. All variables are described in Table 3.17 in the Appendix. Additional details on the benchmark BMA exercise can be found in Table 3.18 and Figure 3.10 in the Appendix.

Thus model averaging techniques corroborate our previous findings concerning publication bias, including less evidence for the bias among IV estimates and estimates for developing countries (these effects are captured by interactions with the standard error). The other three variables found important in all three model averaging techniques are *Developing country*, *IV method*, and *Capital control*. The former two corroborate our results presented in the previous section. A new result is the importance of the control for capital, which is associated with inverse elasticities estimated farther away from zero. Because changes in the capital stock can affect the marginal product of both skilled and unskilled labor, ignoring capital may introduce a bias.

As the bottom line of our analysis we compute an implied elasticity conditional on all collected estimates, our baseline BMA results, and a definition of best practice methodology in the literature. Since best practice is subjective, we choose two distinct strategies. First, we rely on three definitions from the literature: Autor (2014), Card (2009), and Carneiro *et al.* (2022). These are meticulous contributions that have been published in prestigious journals; moreover, they represent the three main streams of the literature using OLS, IV, and natural experiments, respectively. We copy their data and method characteristics and plug those in the values of our variables in order to compute the fitted values from BMA and, hence, the implied (negative inverse) elasticity. Second, we create a subjective definition of best practice based on our reading of the literature.

Our subjective definition of best practice is the following. We plug in zero for the standard error in order to approximately correct for publication bias. We prefer disaggregated panel data and annual granularity. We prefer the multilevel CES structure with all potential control variables included in estimation; furthermore, we prefer dynamic models estimated with unit and time fixed effects and accounting for endogeneity and attenuation bias using instrumental variables. We also prefer studies published in journals with a high impact factor and those with a high number of citations. All other variables (including the ones corresponding to structural variation) are set to their sample means.

Table 3.5: Implied elasticities

	Subjective best practice	Autor (2014)	Card (2009)	Carneiro *et al.* (2022)
All countries	-0.27	-0.13	-0.24	0.05
	(-0.48, -0.05)	(-0.24, -0.02)	(-0.39, -0.09)	(-0.12, 0.23)
	$\sigma = 3.7$	$\sigma = 7.7$	$\sigma = 4.2$	$\sigma = -18.4$
USA	-0.16	-0.02	-0.13	0.16
	(-0.38, 0.06)	(-0.12, 0.07)	(-0.28, 0.02)	(-0.02, 0.34)
	$\sigma = 6.3$	$\sigma = 45.0$	$\sigma = 7.8$	$\sigma = -6.2$
Developing countries	-0.47	-0.33	-0.44	-0.15
	(-0.70, -0.24)	(-0.47, -0.19)	(-0.60, -0.27)	(-0.33, 0.04)
	$\sigma = 2.1$	$\sigma = 3.0$	$\sigma = 2.3$	$\sigma = 6.8$

Notes: The table presents the elasticity of substitution (σ) recovered from the negative inverse elasticity and implied by the results of Bayesian model averaging and i) our definition of best-practice approach, ii) the approach by Autor (2014), iii) the approach by Card (2009), and iv) the approach by Carneiro *et al.* (2022). That is, the table attempts to answer the question what the mean elasticity would look like if the literature was approximately corrected for publication bias and all studies in the literature used the same strategy as the one we prefer or the ones employed by Autor (2014), Card (2009), and Carneiro *et al.* (2022). 95% credible intervals for the negative inverse elasticity are reported in parentheses.

Table 3.5 reports the results. The first row shows the overall estimate, the second row shows the estimate for the US, and the last row shows the estimate for developing countries. Our subjective best practice estimate is in all three cases close to the estimate based on Card (2009). This is because both approaches rely on IV, while OLS and natural experiments in the remaining columns bring inverse elasticities generally close to zero. Our preferred estimate of the implied overall elasticity is 3.7, with the 95% credible interval of $(2, 20)$. The preferred estimate for the US is 6.3; for developing countries it is 2.1. If we ignored any considerations of attenuation bias and instead preferred evidence from natural experiments, we would have to conclude that the implied elasticity is, with the exception of developing countries, close to infinity: a finding even less consistent with the value of 1.5 commonly used for calibrations.

3.4 Conclusion

We collect 682 estimates of the elasticity of substitution between skilled and unskilled labor reported in 77 studies. We measure the extent of two biases that affect the reported inverse elasticity: publication bias (stemming from the underreporting of small estimates) and attenuation bias (stemming from measurement error). Correcting for publication bias slashes the mean negative inverse elasticity from -0.6 to the vicinity of zero, and the result holds when we relax the common meta-analysis assumption of conditional independence of estimates and standard errors. While publication bias corrected estimates stemming from OLS and natural experiments remain close to zero, corrected IV estimates are around -0.25. The result is consistent with attenuation bias in the literature and an implied elasticity of 4 after correction for both biases. The interplay of the two biases in labor economics evokes Griliches (1977), who finds that in measuring the return to education, attenuation bias almost exactly offsets omitted variable bias (which is often correlated with publication bias via specification searching and p-hacking). In our case publication bias dominates attenuation bias.

The aforementioned results hold when we control for additional 24 variables that reflect the context in which the estimates were obtained in the primary studies: for example, variable definition, data characteristics, design of the production function, estimation technique, and publication characteristics. Using so many variables creates model uncertainty problems, and we address them by using both Bayesian model averaging and frequentist model averaging. We find that larger estimated elasticities are associated with data from developed countries and specifications incorporating capital. We then compute the implied elasticity conditional on best practice methodology, based both on prominent studies and our reading of the literature. The implied mean elasticity is again 4, with a 95% credible interval of $(2, 20)$. Because the typical calibration of the elasticity in the literature is 1.5 (Cantore *et al.* 2017), our results suggest that skilled and unskilled labor is substantially more substitutable than commonly thought.

References

Acemoglu, D. (2002): "Technical change, inequality, and the labor market." *Journal of Economic Literature* **40(1)**: pp. 7–72.

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

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models." *Econometrica* **84(6)**: pp. 2155–2182.

Andrews, I. (2018): "Valid Two-Step Identification-Robust Confidence Sets for GMM." *The Review of Economics and Statistics* **100(2)**: pp. 337–348.

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

Andrews, I., J. H. Stock, & L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* **11(1)**: pp. 727–753.

Angrist, J. D. (1995): "The economic returns to schooling in the West Bank and Gaza Strip." *The American Economic Review* **85(5)**: pp. 1065–1087.

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

Askilden, J. E. & O. A. Nilsen (2005): "Apprentices and young workers: A study of the Norwegian youth labour market." *Scottish Journal of Political Economy* **52(1)**: pp. 1–17.

Autor, D. H. (2014): "Skills, education, and the rise of earnings inequality among the 'other 99 percent'." *Science* **344(6186)**: pp. 843–851.

Autor, D. H., L. F. Katz, & M. S. Kearney (2008): "Trends in US wage inequality: Revising the revisionists." *The Review of Economics and Statistics* **90(2)**: pp. 300–323.

Avalos, A. & A. Savvides (2006): "The manufacturing wage inequality in Latin America and East Asia: Openness, technology transfer, and labor supply." *Review of Development Economics* **10(4)**: pp. 553–576.

BAJZIK, J., T. HAVRANEK, Z. IRSOVA, & J. SCHWARZ (2020): "Estimating the Armington elasticity: The importance of study design and publication bias." *Journal of International Economics* **127(C)**. Art. 103383.

BAUM-SNOW, N., M. FREEDMAN, & R. PAVAN (2018): "Why Has Urban Inequality Increased?" *American Economic Journal: Applied Economics* **10(4)**: pp. 1–42.

BEHAR, A. (2010): "The elasticity of substitution between skilled and unskilled labor in developing countries is about 2." *Technical report*, Department of Economics, University of Oxford.

BERGSTROM, V. & E. E. PANAS (1992): "How robust is the capital-skill complementarity hypothesis?" *The Review of Economics and Statistics* **74(3)**: pp. 540–546.

BERMAN, E., J. BOUND, & Z. GRILICHES (1994): "Changes in the Demand for Skilled Labor within U. S. Manufacturing: Evidence from the Annual Survey of Manufactures." *The Quarterly Journal of Economics* **109(2)**: pp. 367–397.

BERNDT, E. R. & L. R. CHRISTENSEN (1974): "Testing for the existence of a consistent aggregate index of labor inputs." *The American Economic Review* **64(3)**: pp. 391–404.

BERNDT, E. R. & C. J. MORRISSON (1979): "Income redistribution and employment effects of rising energy prices." *Resources and Energy* **2(2-3)**: pp. 131–150.

BINELLI, C. (2015): "How the wage-education profile got more convex: Evidence from Mexico." *The B.E. Journal of Macroeconomics* **15(2)**: pp. 509–560.

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

BLANKENAU, W. F. & S. P. CASSOU (2011): "Industry estimates of the elasticity of substitution and the rate of biased technological change between skilled and unskilled labour." *Applied Economics* **43(23)**: pp. 3129–3142.

BLUNDELL, R., D. GREEN, & W. JIN (2016): "The UK wage premium puzzle: How did a large increase in university graduates leave the education premium unchanged?" *IFS Working Papers 1/2016*, The Institute for Fiscal Studies, London.

Boler, E. A. (2016): "Technology-skill complementarity in a globalized world." *Technical report*, Imperial College Business School, Imperial College London, United Kingdom.

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

Borghans, L. & B. ter Weel (2008): "Understanding the technology of computer technology diffusion: Explaining computer adoption patterns and implications for the wage structure." *Journal of Income Distribution* **17(3-4)**: pp. 37–70.

Borjas, G. J. (2003): "The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market." *The Quarterly Journal of Economics* **118(4)**: pp. 1335–1374.

Borjas, G. J., J. Grogger, & G. Hanson (2012): "Comment: On Estimating Elasticities Of Substition." *Journal of the European Economic Association* **10(1)**: pp. 198–210.

Borjas, G. J. & L. F. Katz (2007): "The evolution of the Mexican-born workforce in the United States." In G. J. Borjas (editor), "Mexican immigration to the United States," pp. 13–56. National Bureau of Economic Research Conference Report, University of Chicago Press.

Bound, J., J. Groen, G. Kezdi, & S. Turner (2004): "Trade in university training: Cross-state variation in the production and stock of college-educated labor." *Journal of Econometrics* **121(1)**: pp. 143–173.

Bowles, S. (1970): "Aggregation of labor inputs in the economics of growth and planning: Experiments with a two-level CES function." *Journal of Political Economy* **78(1)**: pp. 68–81.

Bowlus, A. J., L. Lochner, C. Robinson, & E. Suleymanoglu (2022): "Wages, skills, and skill-biased technical change: The canonical model revisited." *Journal of Human Resources* **(forthcoming)**.

Broadstock, D., L. Hunt, & S. Sorrell (2007): "UKERC Review of Evidence for the Rebound Effect - Technical report 3: Elasticity of substitution studies." *UKERC Working Paper Series 11/2007*, UK Energy Research Centre, London.

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

Brodeur, A., M. Le, M. Sangnier, & Y. Zylberberg (2016): "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* **8(1)**: pp. 1–32.

Brown, A. L., T. Imai, F. Vieider, & C. F. Camerer (2022): "Meta-Analysis of Empirical Estimates of Loss-Aversion." *Journal of Economic Literature* **(forthcoming)**.

Brucker, H. & E. J. Jahn (2011): "Migration and Wage-Setting: Reassessing the Labor Market Effects of Migration." *Scandinavian Journal of Economics* **113(2)**: pp. 286–317.

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

Busch, C., D. Krueger, A. Ludwig, I. Popova, & Z. Iftikhar (2020): "Should Germany have built a new wall? Macroeconomic lessons from the 2015-18 refugee wave." *Journal of Monetary Economics* **113(C)**: pp. 28–55.

Caliendo, L., L. Opromolla, F. Parro, & A. Sforza (2021): "Goods and factor market integration: A quantitative assessment of the EU enlargement." *Journal of Political Economy* **129(12)**: pp. 3491–3545.

Cantore, C., F. Ferroni, & M. A. Leon-Ledesma (2017): "The dynamics of hours worked and technology." *Journal of Economic Dynamics and Control* **82(C)**: pp. 67–82.

Card, D. (2009): "Immigration and inequality." *American Economic Review* **99(2)**: pp. 1–21.

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

Card, D. & T. Lemieux (2001): "Can falling supply explain the rising return to college for younger men? A cohort-based analysis." *The Quarterly Journal of Economics* **116(2)**: pp. 705–746.

Carneiro, P., K. Liu, & K. G. Salvanes (2022): "The Supply of Skill and Endogenous Technical Change: Evidence from a College Expansion Reform." *Journal of European Economic Association* **(forthcoming)**.

Carrasco, R., J. Jimeno, & A. Ortega (2015): "Returns to Skills and the Distribution of Wages: Spain 1995–2010." *Oxford Bulletin of Economics and Statistics* **77(4)**: pp. 542–565.

Chirinko, R. S. & D. Mallick (2017): "The Substitution Elasticity, Factor Shares, and the Low-Frequency Panel Model." *American Economic Journal: Macroeconomics* **9(4)**: pp. 225–253.

Choi, K.-S., J.-H. Jeong, & J. Hwa Jung (2005): "The rising supply of college graduates and declining returns for young cohort: The case of Korea." *Global Economic Review* **34(2)**: pp. 167–180.

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

Ciccone, A. & G. Peri (2005): "Long-run substitutability between more and less educated workers: Evidence from US states, 1950–1990." *Review of Economics and Statistics* **87(4)**: pp. 652–663.

Corker, R. & T. Bayoumi (1991): "Apartheid, growth and income distribution in South Africa: Past history and future prospects." *IMF Working Papers 116/1991*, European Department, International Monetary Fund.

Cox, G. & X. Shi (2022): "Simple Adaptive Size-Exact Testing for Full-vector and

Subvector Inference in Moment Inequality Models." *Review of Economic Studies* **(forthcoming)**.

Cruz, M., E. Milet, & M. Olarreaga (2020): "Online exports and the skilled-unskilled wage gap." *PloS ONE* **15(5)**: p. e0232396.

D'Amuri, F., G. I. Ottaviano, & G. Peri (2010): "The labor market impact of immigration in Western Germany in the 1990s." *European Economic Review* **54(4)**: pp. 550–570.

Das, G. G. (1999): "What is assumed in the GTAP database's disaggregation of labor by skill level?" *Monash Economic Working Papers IP-75*, Centre of Policy Studies and Department of Economics, Monash University.

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

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

Denny, M. & M. Fuss (1977): "The use of approximation analysis to test for separability and the existence of consistent aggregates." *The American Economic Review* **67(3)**: pp. 404–418.

Dogan, C. & G. Akay (2019): "The role of labor endowments on industry output in the short run: Evidence from U.S industries." *International Review of Economics & Finance* **60(C)**: pp. 281–291.

Doucouliagos, C. & T. D. Stanley (2013): "Are all economic facts greatly exaggerated? Theory competition and selectivity." *Journal of Economic Surveys* **27(2)**: pp. 316–339.

Dougherty, C. R. (1972): "Estimates of labor aggregation functions." *Journal of Political Economy* **80(6)**: pp. 1101–1119.

Dupuy, A. (2007): "Will the skill-premium in the Netherlands rise in the next decades?" *Applied Economics* **39(21)**: pp. 2723–2731.

Dupuy, A. & P. Marey (2008): "Shifts and twists in the relative productivity of skilled labor." *Journal of Macroeconomics* **30(2)**: pp. 718–735.

Dustmann, C., J. Ludsteck, & U. Schonberg (2009): "Revisiting the German Wage Structure." *The Quarterly Journal of Economics* **124(2)**: pp. 843–881.

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

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

Fallon, P. R. & P. R. G. Layard (1975): "Capital-skill complementarity, income distribution, and output accounting." *Journal of Political Economy* **83(2)**: pp. 279–302.

Feldkircher, M. & S. Zeugner (2012): "The impact of data revisions on the robustness of growth determinants—a note on determinants of economic growth: Will data tell?" *Journal of Applied Econometrics* **27(4)**: pp. 686–694.

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

Fernandez, M. & J. Messina (2018): "Skill premium, labor supply, and changes in the structure of wages in Latin America." *Journal of Development Economics* **135(C)**: pp. 555–573.

Fitzgerald, J. D. & I. Kearney (2000): "Convergence in living standards in Ireland: The role of the new economy." *ESRI Working Paper Series 134/2000*, Economic and Social Research Institute.

Foldvari, P. & B. van Leeuwen (2006): "Is there an optimal allocation of skills? A two-level CES approach." *Technical report*, Department of Economics, University of Warwick.

Freeman, R. B. (1975): "Overinvestment in college training?" *Journal of Human Resources* **10(3)**: pp. 287–311.

Freeman, R. B. & J. L. Medoff (1982): "Substitution between production labor and other Inputs in unionzed and nonunionized manufacturing." *The Review of Economics and Statistics* **64(2)**: pp. 220–233.

Furukawa, C. (2020): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Working paper*, MIT.

Gallego, F. A. (2012): "Skill premium in Chile: Studying skill upgrading in the South." *World Development* **40(3)**: pp. 594–609.

Gancia, G. A., A. Muller, & F. Zilibotti (2013): "Structural development accounting." In D. Acemoglu, M. Arellano, & E. Dekel (editors), "Advances in Economics and Econometrics: Tenth World Congress," volume 2 of *Econometric Society Monographs*, chapter 9, pp. 373–418. Cambridge: Cambridge University Press.

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

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." In "IMS Collections Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown," volume 6, p. 158–165. Institute of Mathematical Statistics.

Gerber, A. & N. Malhotra (2008): "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* **3(3)**: pp. 313–326.

Giannarakis, P. (2017): "Can job displacements explain the UK's productivity puzzle?" *Technical report*, University of Southampton.

Glitz, A. & D. Wissmann (2021): "Skill premiums and the supply of young workers in Germany." *Labour Economics* **72(C)**. Art. 102034.

Goldin, C. D. & L. F. Katz (2009): *The race between education and technology.* Cambridge, MA: Harvard University Press, 1st edition.

Griliches, Z. (1977): "Estimating the Returns to Schooling: Some Econometric

Problems." *Econometrica* **45(1)**: pp. 1–22.

Gunadi, C. (2019): "An inquiry on the impact of highly-skilled STEM immigration on the U.S. economy." *Labour Economics* **61(C)**. Art. 101751.

Gyimah-Brempong, K. & A. O. Gyapong (1992): "Elasticities of factor substitution in the production of education." *Economics of Education Review* **11(3)**: pp. 205–217.

Hamermesh, D. S. (1996): *Labor demand.* Princeton, NJ: Princeton University Press.

Hansen, B. E. (2007): "Least Squares Model Averaging." *Econometrica* **75(4)**: pp. 1175–1189.

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

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

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

Heckman, J. J., L. Lochner, & C. Taber (1998): "Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents." *Review of Economic Dynamics* **1(1)**: pp. 1–58.

Hendricks, L. & T. Schoellman (2018): "Human Capital and Development Accounting: New Evidence from Wage Gains at Migration." *The Quarterly Journal of Economics* **133(2)**: pp. 665–700.

Hendricks, L. & T. Schoellman (2022): "Skilled Labor Productivity and Cross-country Income Differences." *American Economic Journal: Macroeconomics*

(forthcoming).

Hijzen, A., H. Gorg, & R. Hine (2005): "International Outsourcing and the Skill Structure of Labour Demand in the United Kingdom." *The Economic Journal* **115(506)**: pp. 860–878.

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

Ioannidis, J. P. A., T. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *The Economic Journal* **127(605)**: pp. F236–F265.

Iwasaki, I. (2022): "The finance-growth nexus in Latin America and the Caribbean: A meta-analytic perspective." *World Development* **149(C)**. Art. 105692.

Jamet, S. (2005): "From sectoral to macro-economic impact of cuts in social contributions (in French: De l'impact sectoriel a l'effet macro-economique des allegements de cotisations sociales)." *Revue Francaise D'Economie* **19(3)**: pp. 57–90.

Jeffreys, H. (1961): *Theory of Probability.* Oxford Classic Texts in the Physical Sciences. Oxford: Oxford University Press, third edition.

Jensen, G. A. & M. A. Morrisey (1986): "The role of physicians in hospital production." *The Review of Economics and Statistics* **68(3)**: pp. 432–442.

Jerzmanowski, M. & R. Tamura (2020): "Aggregate Elasticity of Substitution between Skills: Estimates from a Macroeconomic Approach." *MPRA Paper 100768*, University Library of Munich, Germany.

Johnson, G. E. (1970): "The demand for labor by educational category." *Southern Economic Journal* **37(2)**: pp. 190–204.

Johnson, M. & M. P. Keane (2013): "A Dynamic Equilibrium Model of the US Wage Structure, 1968-1996." *Journal of Labor Economics* **31(1)**: pp. 1–49.

Katz, L. F. & K. M. Murphy (1992): "Changes in relative wages, 1963–1987: Supply and demand factors." *The Quarterly Journal of Economics* **107(1)**: pp. 35–78.

KAWAGUCHI, D. & Y. MORI (2016): "Why has wage inequality evolved so differently between Japan and the US? The role of the supply of college-educated workers." *Economics of Education Review* **52(1)**: pp. 29–50.

KEARNEY, I. (1997): "Estimating the demand for skilled labour, unskilled labour and clerical workers: A dynamic framework." *ESRI Working Paper 91*, The Economic and Social Research Institute.

KESSELMAN, J. R., S. H. WILLIAMSON, & E. R. BERNDT (1977): "Tax credits for employment rather than investment." *The American Economic Review* **67(3)**: pp. 339–349.

KIM, D. I. (2005): "Growth in college education and wage differentials in Korea." *Seoul Journal of Economics* **18(2)**: pp. 87–123.

KIYOTA, K. & Y. KUROKAWA (2019): "Intermediate Goods-Skill Complementarity." *Keio-IES Discussion Paper Series 2019-013*, Institute for Economics Studies, Keio University.

KLENOW, P. J. & A. RODRIGUEZ-CLARE (1997): "The neoclassical revival in growth economics: Has it gone too far?" *NBER Macroeconomics Annual* **12**: pp. 73–114.

KLOTZ, B., R. MADOO, & R. HANSEN (1980): "A study of high and low "labor productivity" establishments in US manufacturing." In J. W. KENDRICK & B. N. VACCARA (editors), "New developments in productivity measurement," chapter 3, pp. 239–292. University of Chicago Press.

KRANZ, S. & P. PUTZ (2022): "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* **(forthcoming)**.

KRUSELL, P., L. E. OHANIAN, J.-V. RIOS-RULL, & G. L. VIOLANTE (2000): "Capital-skill complementarity and inequality: A macroeconomic analysis." *Econometrica* **68(5)**: pp. 1029–1053.

KWACK, S. Y. (2012): "Wage Inequality and the Contribution of Capital, Differential Labor Quality and Efficiency to Economic Growth in Korea, 1965-2007." *Seoul*

Journal of Economics **25(1)**: pp. 1–23.

Ley, E. & M. F. J. Steel (2009): "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* **24(4)**: pp. 651–674.

Li, B. (2010): "Multinational production and choice of technologies: New evidence on skill-biased technological change from China." *Economics Letters* **108(2)**: pp. 181–183.

Lindquist, M. J. (2005): "Capital-skill complementarity and inequality in Sweden." *Scandinavian Journal of Economics* **107(4)**: pp. 711–735.

Malmberg, H. (2018): "How does the efficiency of skilled labor vary across rich and poor countries? An analysis using trade and industry data." *Working paper*, Stanford Institute for Economic Policy Research, Stanford University.

Manacorda, M., A. Manning, & J. Wadsworth (2012): "The Impact Of Immigration On The Structure Of Wages: Theory And Evidence From Britain." *Journal of the European Economic Association* **10(1)**: pp. 120–151.

Manacorda, M., C. Sanchez-Paramo, & N. Schady (2010): "Changes in returns to education in Latin America: The role of demand and supply of skills." *Industrial & Labor Relations Review* **63(2)**: pp. 307–326.

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

McAdam, P. & A. Willman (2018): "Unraveling The Skill Premium." *Macroeconomic Dynamics* **22(1)**: pp. 33–62.

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

Medina, C. & C. M. Posso (2010): "Technical change and polarization of the labor market: Evidence for Brazil, Colombia and Mexico." *Working Paper Series*

"Borradores de Economia" 614, Banco de la Republica de Colombia.

Mello, M. (2011): "Another look at panel estimates of the elasticity of substitution." *Working paper*, IBMEC Business School, Rio de Janeiro.

Mollick, A. V. (2008): "The rise of the skill premium in Mexican maquiladoras." *The Journal of Development Studies* **44(9)**: pp. 1382–1404.

Murphy, K. M., W. C. Riddell, & P. M. Romer (1998): "Wages, skills, and technology in the United States and Canada." *NBER Working Paper 6638*, National Bureau of Economic Research.

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

Nissim, J. (1984): "The price responsiveness of the demand for labour by skill: British mechanical engineering: 1963-1978." *The Economic Journal* **94(376)**: pp. 812–825.

Ohanian, L., M. Orak, & S. Shen (2021): "Revisiting Capital-Skill Complementarity, Inequality, and Labor Share." *International Finance Discussion Papers 1319*, Board of Governors of the Federal Reserve System (U.S.).

Ottaviano, G. & G. Peri (2012): "Rethinking The Effect Of Immigration On Wages." *Journal of the European Economic Association* **10(1)**: pp. 152–197.

Psacharopoulos, G. & K. Hinchliffe (1972): "Further evidence on the elasticity of substitution among different types of educated labor." *Journal of Political Economy* **80(4)**: pp. 786–792.

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

Razzak, W. A. & J. C. Timmins (2008): "A macroeconomic perspective on skill shortages and the skill premium in New Zealand." *Australian Economic Papers* **47(1)**: pp. 74–91.

Reijnders, L., M. Timmer, & X. Ye (2021): "Labour demand in global value chains: Is there a bias against unskilled work?" *The World Economy* **44(9)**: pp. 2547–2571.

Reshef, A. (2007): "Heckscher-Ohlin and the global increase of skill premia: Factor intensity reversals to the rescue." *Working paper*, New York University.

Riano, A. (2009): "Trade, technology adoption and the rise of the skill premium in Mexico." *Working paper*, Globalisation and Economic Policy Centre, University of Nottingham.

Robbins, D. J. (1996): "Evidence on trade and wages in the developing world." *Working Papers 119*, Organisation for Economic Co-operation and Development.

Roodman, D., J. G. MacKinnon, M. O. Nielsen, & M. D. Webb (2018): "Fast and wild: Bootstrap inference in Stata using boottest." *Working Paper 1406*, Department of Economics, Queen's University, Canada: Kingston.

Silva, J. C. (2008): "Young employment, job-skill composition and minimum wages: Evidence from a "natural experiment"." *NIPE Working paper 4/2008*, Universidade do Minho, Portugal.

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

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

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

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

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

Stanley, T. D., H. Doucouliagos, J. P. A. Ioannidis, & E. C. Carter (2021):

"Detecting publication selection bias through excess statistical significance." *Research Synthesis Methods* **12(6)**: pp. 776–795.

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

Sun, L. (2018): "Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models." *Stata Journal* **18(4)**: pp. 803–825.

Tinbergen, J. (1974): "Substitution of graduate by other labour." *Kyklos* **27(2)**: pp. 217–226.

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

te Velde, D. W. & O. Morrissey (2004): "Foreign direct investment, skills and wage inequality in East Asia." *Journal of the Asia Pacific Economy* **9(3)**: pp. 348–369.

Verdugo, G. (2014): "The great compression of the French wage structure, 1969–2008." *Labour Economics* **28(C)**: pp. 131–144.

Wei, X., G. Onel, Z. Guan, & F. M. Roka (2019): "Substitution between immigrant and native farmworkers in the United States: Does legal status matter?" *IZA Journal of Development and Migration* **10(1)**: pp. 1–27.

Welch, F. (1970): "Education in production." *Journal of Political Economy* **78(1)**: pp. 35–59.

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

Yang, D. (2012): "The responsiveness of firms to labor market changes in China: Evidence from micro level data." *Working Paper 6239*, China Institute for Employment Research and IZA Institute of Labor Economics.

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

68(4): pp. 1–37.

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

3.A The Elasticity Dataset

The elasticity of substitution between skilled and unskilled labor is usually defined as the change of the ratio in which these two factors are used in production divided by the change of the ratio of their marginal products. Under perfect competition, production factors are paid their marginal products and the elasticity can be written as

$$\sigma_{US} = \frac{\frac{d(L_U/L_S)}{L_U/L_S}}{\frac{d(w_S/w_U)}{w_S/w_U}} = -\frac{d\ log(L_U/L_S)}{d\ log(w_U/w_S)}, \tag{3.1}$$

where L_S and L_U denote skilled and unskilled labor; w_S and w_U denote their respective wage rates. Under a quasi-concave production function the elasticity of substitution attains any value from zero to infinity. If $\sigma = 0$, the two types of labor form perfect complements. Fixed proportions of the two inputs are needed to increase production; they cannot be substituted for each other. If $\sigma \in (0, 1)$, skilled and unskilled workers are gross complements: an increased supply of skilled workers increases the demand for unskilled workers. A unitary elasticity implies that relative quantity changes are exactly proportional to relative price changes. If $\sigma > 1$, skilled and unskilled workers form gross substitutes: unskilled workers can more easily work in positions intended for skilled workers (though with a lower productivity), and skilled workers can be tapped for a menial job. An increased supply of skilled workers decreases the demand for unskilled workers. Many researchers estimating the elasticity start with the following constant elasticity of substitution (CES) production function:

$$Y = [\alpha(aL_S)^\rho + (1-\alpha)(bL_U)^\rho]^{\frac{1}{\rho}}, \tag{3.2}$$

where skilled labor L_S and unskilled labor L_U are the sole factors of production, a and b are indices of factor-augmenting technology, and α is a technology parameter interpretable as indexing the "share of work" allocated to L_S. The elasticity can be derived from the parameter ρ as $\sigma = \frac{1}{1-\rho}$.

Whether researchers assume a one-level CES function or a nested one (also taking into account other inputs, such as capital), they typically employ the following steps. First, marginal products are obtained by taking derivatives of Y with respect to L_S and L_U. The assumption of competitive labor markets implies the equality of the wage ratio and the ratio of marginal products. Substituting $(\sigma - 1)/\sigma$ for ρ then leads to the definition of the skill premium $\frac{w_S}{w_U}$:

$$\frac{w_S}{w_U} = \frac{\alpha}{1-\alpha}\left(\frac{a}{b}\right)^{\frac{\sigma-1}{\sigma}}\left(\frac{L_S}{L_U}\right)^{-\frac{1}{\sigma}}. \tag{3.3}$$

Taking logarithms produces a specification that can be estimated:

$$\ln\left(\frac{w_S}{w_U}\right) = \ln\left(\frac{\alpha}{1-\alpha}\right) + \frac{\sigma-1}{\sigma}\ln\left(\frac{a}{b}\right) - \frac{1}{\sigma}\ln\left(\frac{L_S}{L_U}\right). \tag{3.4}$$

Table 3.6: The studies used in the meta-analysis (reporting both direct and inverse estimates)

Acemoglu (2002)	D'Amuri *et al.* (2010)	Kesselman *et al.* (1977)
Angrist (1995)	Das (1999)	Kim (2005)
Askilden & Nilsen (2005)	Denny & Fuss (1977)	Kiyota & Kurokawa (2019)
Autor (2014)	Dogan & Akay (2019)	Klenow & Rodriguez-Clare (1997)
Autor *et al.* (2008)	Dougherty (1972)	Klotz *et al.* (1980)
Avalos & Savvides (2006)	Dupuy & Marey (2008)	Krusell *et al.* (2000)
Baum-Snow *et al.* (2018)	Dupuy (2007)	Kwack (2012)
Behar (2010)	Dustmann *et al.* (2009)	Li (2010)
Bergstrom & Panas (1992)	Fallon & Layard (1975)	Lindquist (2005)
Berndt & Christensen (1974)	Fernandez & Messina (2018)	Malmberg (2018)
Berndt & Morrisson (1979)	Fitzgerald & Kearney (2000)	Manacorda *et al.* (2010)
Binelli (2015)	Foldvari & van Leeuwen (2006)	Manacorda *et al.* (2012)
Blankenau & Cassou (2011)	Freeman & Medoff (1982)	McAdam & Willman (2018)
Blundell *et al.* (2016)	Freeman (1975)	Medina & Posso (2010)
Boler (2016)	Gallego (2012)	Mello (2011)
Borghans & ter Weel (2008)	Gancia *et al.* (2013)	Mollick (2008)
Borjas & Katz (2007)	Giannarakis (2017)	Murphy *et al.* (1998)
Borjas (2003)	Glitz & Wissmann (2021)	Nissim (1984)
Borjas *et al.* (2012)	Goldin & Katz (2009)	Ohanian *et al.* (2021)
Bound *et al.* (2004)	Gunadi (2019)	Ottaviano & Peri (2012)
Bowles (1970)	Gyimah-Brempong & Gyapong (1992)	Psacharopoulos & Hinchliffe (1972)
Bowlus *et al.* (2022)	Heckman *et al.* (1998)	Razzak & Timmins (2008)
Brucker & Jahn (2011)	Hendricks & Schoellman (2018)	Reijnders *et al.* (2021)
Busch *et al.* (2020)	Hendricks & Schoellman (2022)	Reshef (2007)
Caliendo *et al.* (2021)	Hijzen *et al.* (2005)	Riano (2009)
Card & Lemieux (2001)	Jamet (2005)	Robbins (1996)
Card (2009)	Jensen & Morrisey (1986)	Silva (2008)
Carneiro *et al.* (2022)	Jerzmanowski & Tamura (2020)	Tinbergen (1974)
Carrasco *et al.* (2015)	Johnson & Keane (2013)	te Velde & Morrissey (2004)
Choi *et al.* (2005)	Johnson (1970)	Verdugo (2014)
Ciccone & Peri (2005)	Katz & Murphy (1992)	Wei *et al.* (2019)
Corker & Bayoumi (1991)	Kawaguchi & Mori (2016)	Welch (1970)
Cruz *et al.* (2020)	Kearney (1997)	Yang (2012)

The main coefficient of interest, the negative inverse of the elasticity $(-1/\sigma)$, can thus be interpreted as the causal effect of the relative supply of skilled labor on the wage premium to skills (in percentage terms). The term capturing skill-biased technical change (a/b), and thus demand for skills, is usually proxied by a time trend. Some authors estimate the regression in a reversed form (regressing relative skill supply on the skill premium), which gives them a direct estimate of the elasticity, not its inverse (see details in Section 3.B). Details on various specification and estimation techniques employed in the literature are available in Section 3.D. In data collection and reporting we follow the guidelines compiled by the Meta-Analysis in Economics Research Network (Havranek *et al.* 2020).

We search for studies in Google Scholar, which allows our search query to go through the full text of research papers, not just the title, keyword, and abstract, which is the case for most other databases. We examine the first 500 studies returned by the search. We read the abstract of each study to identify those that may potentially include empirical estimates of the elasticity; we then download such studies and read them in detail. Furthermore, we inspect the lists of references of all these studies to find any potentially important papers omitted by our Google Scholar search; we terminate the literature search on October 31, 2021. The data and code are available at `meta-analysis.cz/skill`.

Three co-authors have collected 1/3 of the data each and randomly checked 20% of the data collected by the remaining two co-authors in order to identify and correct potential inconsistencies in coding. The final sample includes 1,096 estimates of the elasticity collected from 99 studies listed in Table 3.6; we call them primary studies. (Note that only 965 of these estimates are reported together with a standard error, which means that we can use them in publication bias tests.) The oldest study was published in 1970, the most recent is forthcoming as of 2022, covering five decades of research. To give less weight to outliers we winsorize estimates at the 1% level. The histogram of the collected estimates is presented in Figure 3.5 for both inverse elasticities (from regressions of the skill premium on relative labor supply) and direct elasticities (from reverse regressions and translog specifications, see Section 3.B for details). The literature uses both streams of studies for calibration, but since in most situations causality can be expected to run from labor supply changes to changes in the skill premium, interpretation of the directly estimated elasticities is less clear. We collect data from both groups of studies but focus on inverse elasticities in the main analysis. From the histogram we observe that the values of the elasticity that are used for calibration (often between 1 and 2) form a minority of empirical estimates in both groups of studies.

The mean elasticity, directly estimated, is 0.9 (averaged over 414 estimates). The mean estimated negative inverse elasticity is -0.6 (averaged over 682 estimates), which implies an elasticity of 1.8. Summary statistics for various subsamples of the literature are available in Table 3.7. Figure 3.6 shows that estimates differ substantially not only across but also within individual studies. Substantial heterogeneity is one stylized fact of the literature on skill substitution. The second stylized fact that arises from a bird's-eye view of the data is the break in the frequency of estimated negative inverse elasticities at -1 and especially at 0 (Figure 3.5), which might suggest publication bias.

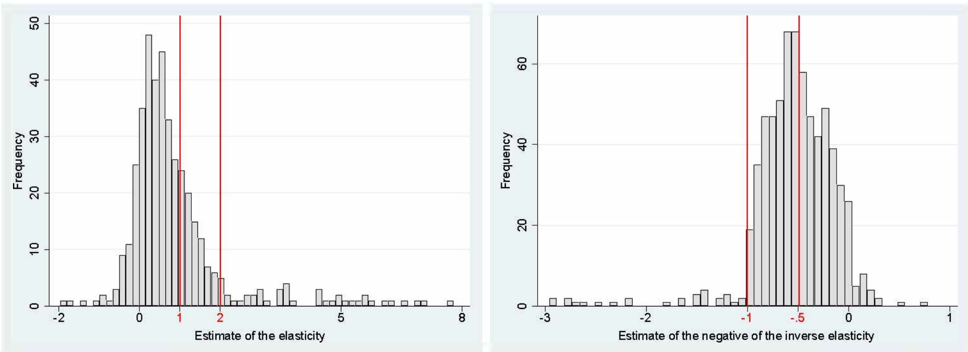
Table 3.7: Summary statistics for different subsets of the literature

		Unweighted			Weighted		
	No. of obs.	Mean	95% conf. int.		Mean	95% conf. int.	
Direct elasticity estimate	414	0.921	0.788	1.054	1.507	1.343	1.671
Negative inverse elasticity	682	-0.554	-0.589	-0.520	-0.545	-0.582	-0.508
Subsamples of the negative inverse elasticity							
Data characteristics							
Higher frequency	65	-0.263	-0.319	-0.207	-0.277	-0.332	-0.222
Annual frequency	528	-0.576	-0.612	-0.540	-0.564	-0.603	-0.526
Lower frequency	89	-0.639	-0.785	-0.493	-0.893	-1.088	-0.699
Micro data	179	-0.584	-0.655	-0.513	-0.456	-0.519	-0.393
Sectoral data	112	-0.734	-0.863	-0.606	-0.872	-1.027	-0.718
Aggregated data	391	-0.489	-0.523	-0.455	-0.510	-0.548	-0.473
Cross-section	58	-0.635	-0.794	-0.476	-0.631	-0.782	-0.481
Panel or time-series	624	-0.547	-0.582	-0.512	-0.529	-0.566	-0.491
Structural variation							
United States	285	-0.416	-0.452	-0.380	-0.492	-0.532	-0.452
Developed country	443	-0.528	-0.577	-0.479	-0.547	-0.597	-0.497
Developing country	152	-0.474	-0.511	-0.437	-0.476	-0.519	-0.433
Manufacturing sector	95	-0.821	-0.924	-0.717	-0.675	-0.805	-0.545
Regional estimate	93	-0.449	-0.496	-0.402	-0.384	-0.435	-0.332
Country estimate	583	-0.579	-0.618	-0.539	-0.570	-0.612	-0.529
Design of the production function							
One-level CES function	229	-0.465	-0.512	-0.419	-0.479	-0.529	-0.428
Multilevel CES function	453	-0.599	-0.646	-0.553	-0.588	-0.638	-0.537
Estimation technique							
Dynamic model	52	-0.574	-0.717	-0.431	-0.662	-0.828	-0.496
Unit fixed effects	342	-0.669	-0.723	-0.615	-0.637	-0.699	-0.576
Time fixed effects	157	-0.504	-0.564	-0.443	-0.414	-0.472	-0.355
OLS method	362	-0.472	-0.512	-0.432	-0.531	-0.577	-0.484
IV method	264	-0.584	-0.621	-0.547	-0.514	-0.555	-0.473
Natural experiment	40	-1.191	-1.535	-0.847	-0.927	-1.235	-0.618
Publication characteristics							
Unpublished study	202	-0.778	-0.857	-0.699	-0.790	-0.882	-0.698
Published study	480	-0.460	-0.493	-0.427	-0.484	-0.522	-0.447
Top journal publication	131	-0.442	-0.492	-0.393	-0.410	-0.458	-0.362

Notes: The table reports summary statistics of the reported elasticity of substitution for different subsets of the literature and includes also estimates reported without the standard error (which are excluded in the analysis of publication bias). The exact definition of the variables is available in Table 2.4. Weighted = estimates are weighted by the inverse of the number of estimates reported per study.

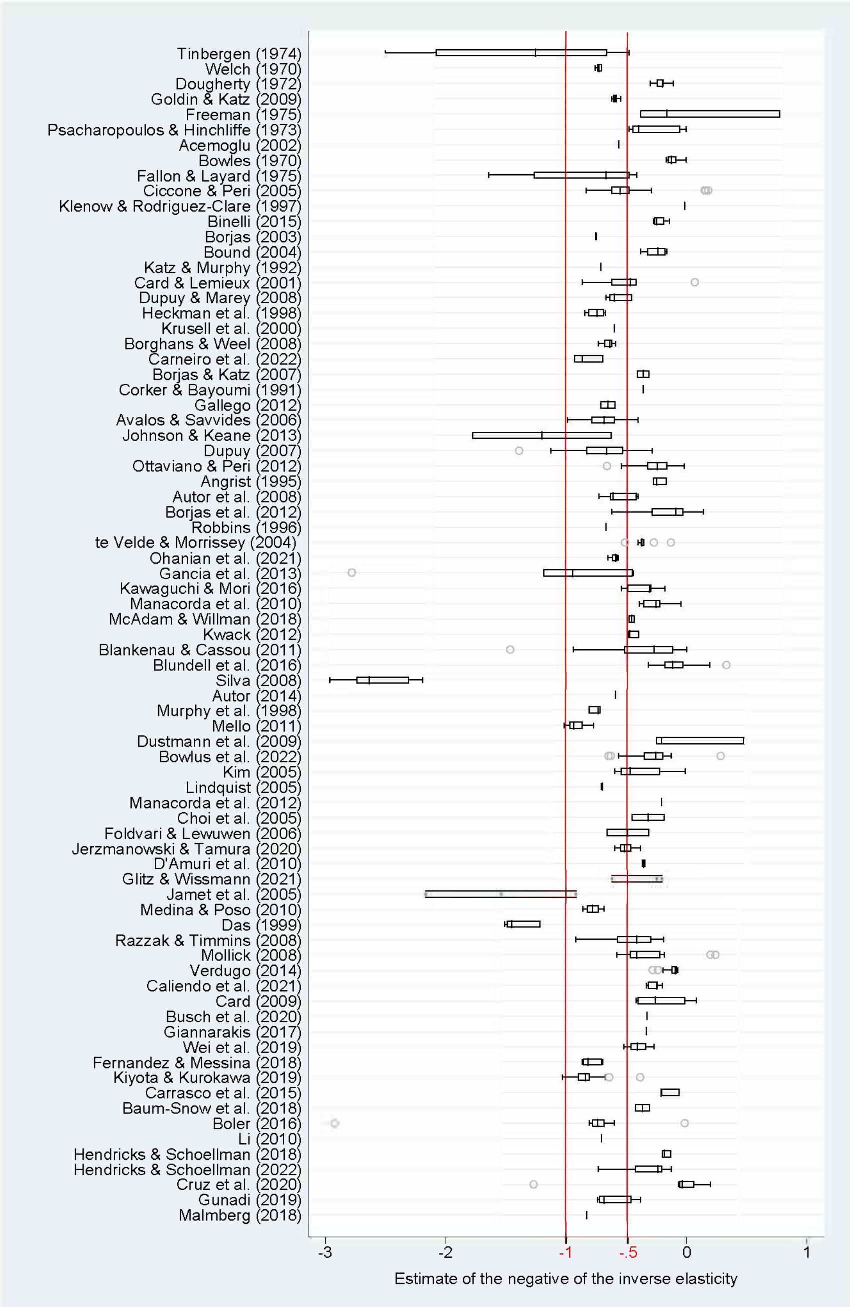
Figure 3.5: Distribution of the reported estimates

(a) Direct elasticity (b) Negative inverse elasticity



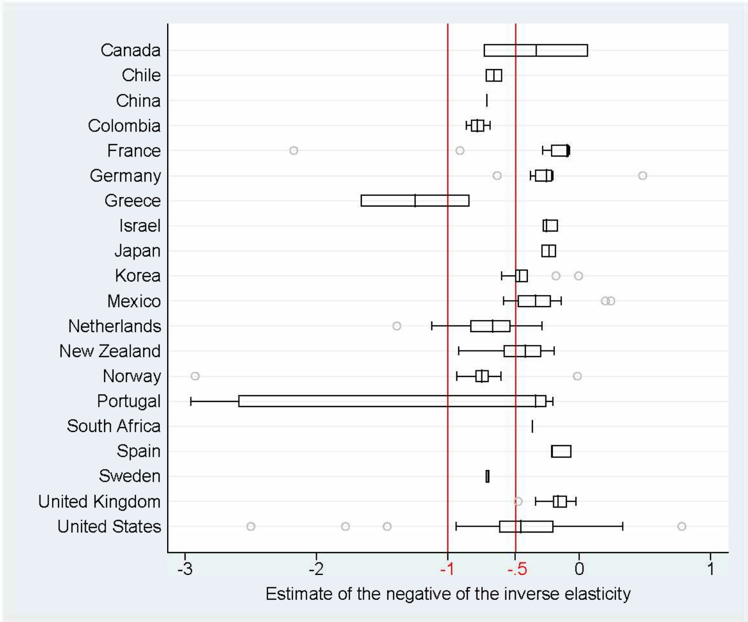
Notes: The figure depicts a histogram of the elasticicities reported by individual studies. The vertical lines denote the interval for the elasticity of $\langle 1, 2 \rangle$, from which most of the values used for calibrations are drawn. For ease of exposition, outliers are excluded from the figure but included in all statistical tests.

Figure 3.6: Estimates of the negative inverse elasticity vary both within and across studies



Notes: The studies are sorted by the age of the data they use from oldest to youngest. 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. The vertical lines denote the interval for the elasticity of $\langle 1, 2 \rangle$, from which most of the values used for calibrations are drawn. For ease of exposition, outliers are excluded from the figure but included in all statistical tests.

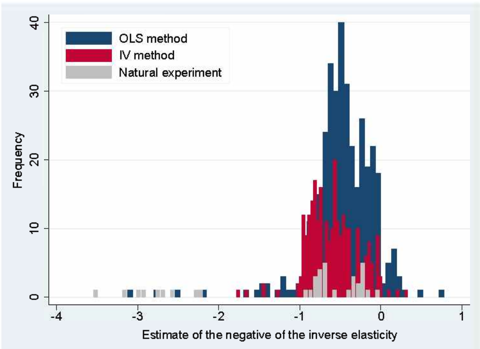
Figure 3.7: Cross-country heterogeneity in the negative inverse elasticity



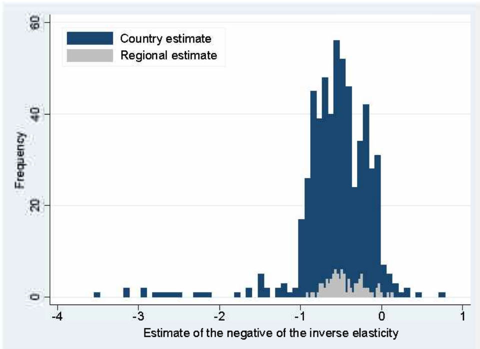
Notes: The length of each box represents the interquartile range (P25-P75), and the dividing line inside the box is the median value. The whiskers represent the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. The vertical lines denote the interval for the elasticity of $\langle 1, 2\rangle$, from which most of the values used for calibrations are drawn. For ease of exposition, outliers are excluded from the figure but included in all statistical tests.

Figure 3.8: Prima facie patterns in the reported negative inverse elasticities

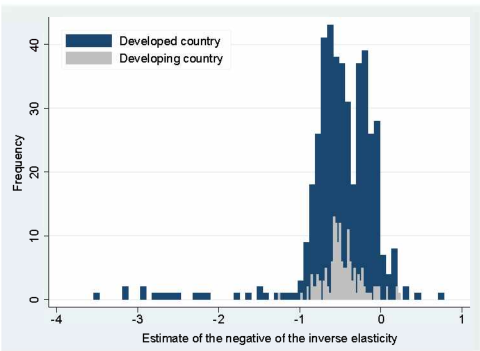
(a) Estimation approach



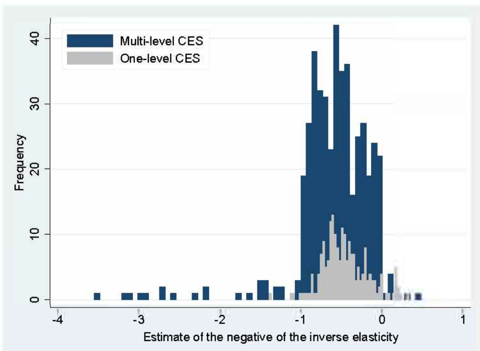
(b) Estimate aggregation



(c) Country classification



(d) Production function



Notes: We use the IMF definition to classify countries as developed or developing.

3.B Direct Estimates of the Elasticity

As we have noted, most studies identify the elasticity of substitution between skilled and unskilled labor by regressing relative wages on the relative labor supply; the estimated coefficient represents the (negative) inverse elasticity. This approach is intuitive because relative labor supply cannot change fast in reaction to changes in relative wages. If a study is based on genuine random variation in the relative labor supply, then it can identify the causal effect. Other studies estimate the elasticity directly by either running a reverse regression or using a translog specification. These two groups of studies cannot be combined in an analysis of publication bias because the inversion necessary for such a combination creates a mechanical relationship between estimates and standard errors. Therefore the two groups of regression coefficients have to be analyzed separately. In the main body of the paper we focus on the inverse elasticity for three reasons.

First, out of the 99 studies that we find for this meta-analysis, only 24 studies estimate the elasticity directly (including two studies that also report inverse estimates). Second, only 75% of the direct estimates are reported together with standard errors, compared to 96% of the inverse estimates. This observation, especially in the case of studies using the translog function, further decreases the power of publication bias tests for direct estimates. Table 3.8 shows that studies reporting direct elasticities are also typically older and published in outlets with a smaller impact factor. Third, in most cases we find the identification arguments for reverse regressions less persuasive: it is unclear what can be achieved by regressing the “treatment” (relative labor supply) on the “outcome” (relative wages).

Table 3.8: Studies relying on direct estimates look worse on paper

Reported standard errors	-0.771**
	(0.377)
Publication year	-0.530**
	(0.224)
Impact factor	-0.393**
	(0.195)
Citations	-0.207
	(0.174)
Constant	2.219**
	(0.887)
Observations (studies)	99

Notes: The table shows the results of a probit regression (response variable = 1 if the study reports only direct estimates of the elasticity). Reported standard errors = 1 if the study reports standard errors, p-values, or t-statistics for any of its point estimates, Publication year in logs, Impact factor is the discounted recursive RePEc impact factor of the outlet, Citations = log of the number of per-year citations of the study in Google Scholar. Robust standard errors in parentheses.

Table 3.9 lists all the studies that report direct estimates of the elasticity of substitution between skilled and unskilled labor. We divide them into five categories according to the identification arguments that they use: studies using highly disaggregated data, studies introducing a new theory, studies concerned about measurement error, studies using the translog production function, and studies using the translog cost function. Bowles (1970) and Ciccone & Peri (2005) also report, in fact as their key results, estimates of inverse elasticities, so we include those in the main analysis presented in the paper.

Regarding disaggregated data, five studies make, explicitly or implicitly, the assumption that their dataset is disaggregated enough to allow them analyze firm-level demand for skills. Under perfect competition, the relative price of skill is assumed to be fixed. A persuasive dataset and identification strategy are provided by Boler (2016) for the 2002 R&D expenditures reform in Norway, but for the remaining studies (three of them unpublished, the remaining one published in 1982) identification is less clear.

Regarding studies presenting a new theory, Brucker & Jahn (2011) assume monopolistic competition and wage-setting framework, in which wages are fixed first and labor outcomes are determined later. The unpublished paper by Behar (2010) features a standard regression of the wage premium on the relative relative labor supply but the author interprets the regression coefficient as a direct estimate of the elasticity based on a theoretical framework that incorporates technology import effects for developing countries.

Regarding measurement error, Bowles (1970) estimates a reverse regression to establish an upper bound for the extent of attenuation bias but otherwise focuses on a standard regression of the wage premium on the relative labor supply. The motivation for the use of reverse regression by Johnson (1970) is less clear but implicitly seems to be based on the idea that the wage premium is measured with less noise.

Regarding the translog production function, its use is motivated by the desire to relax the assumption that the elasticity is constant along the demand curve. Researchers regress skilled workers' share of wages on the relative labor supply; a more detailed description is available in the informative paper by Ciccone & Peri (2005, pp. 659–661), who also estimate the inverse elasticity. Hence the identification argument for the translog production function is persuasive, at least in the case of Ciccone & Peri (2005), and generally similar to the standard case for inverse estimates—but here the regression coefficient does not need to be inverted to yield the elasticity and thus cannot be pooled with the dataset used in the main body of our paper. The subgroup is unfortunately too small to be analyzed separately.

Table 3.9: The 24 studies reporting direct estimates of the elasticity

Category	Study	Justification
Disaggregated	Boler (2016) Freeman & Medoff (1982) Reshef (2007) Riano (2009) Yang (2012)	Estimated for individual firms or small industries, authors assume that the relative price of skilled labor is fixed. Reshef (2007, p. 13): "So it is not a terrible sin to ignore the classic simultaneous demand-supply identification problem for an individual industry."
Theory	Behar (2010) Brucker & Jahn (2011)	The identification approach relies on a concept different from the rest of the literature (monopolistic competition, wage setting, technology import effects). Brucker & Jahn (2011, p. 302): "It follows from our wage-setting framework that labor demand is endogenously determined once wages are fixed."
Measurement error	Bowles (1970) Johnson (1970)	Estimated in response to concerns regarding attenuation bias. Bowles (1970, p. 75): "An upper limit to this bias can be established by making the ratio of labor quantities the dependent variable."
Translog production	Berndt & Christensen (1974) Ciccone & Peri (2005) Denny & Fuss (1977) Jensen & Morrisey (1986)	Relaxing the assumption that the elasticity is constant along the demand curve; authors regress skilled workers' share of wages on relative labor supply. Ciccone & Peri (2005, p. 659): "The key parameter can be estimated consistently using the same instruments and the same identifying assumptions as in the CES case."
Translog cost	Askilden & Nilsen (2005) Bergstrom & Panas (1992) Berndt & Morrisson (1979) Dogan & Akay (2019) Fitzgerald & Kearney (2000) Gyimah-Brempong & Gyapong (1992) Hijzen *et al.* (2005) Kearney (1997) Kesselman *et al.* (1977) Klotz *et al.* (1980) Nissim (1984)	Relaxing the assumption that the elasticity is constant; authors regress skilled workers' share of wages on relative wages (following Shephard's duality theory). Often used with relatively disaggregated data, but we have found no explicit justification of this identification approach in the specific context of skill substitution. In addition, standard errors are rarely reported in this group of studies.

Notes: The table lists studies that report direct estimates of the elasticity of substitution between skilled and unskilled labor—in contrast with the estimates used in our main analysis, which need to be inverted to yield the elasticity (the standard approach in the literature). Because inversion creates a mechanical relationship between estimates and standard errors, these two groups of studies cannot be pooled together in an analysis of publication bias. Note that Bowles (1970) and Ciccone & Peri (2005) also report, in fact as their main results, inverse elasticities, so we include those in the main analysis.

Regarding the translog cost function, this is the most common approach for the direct estimation of the elasticity of substitution: 11 studies in our dataset use this technique. The motivation is similar to the one for the translog production function, but here researchers regress skilled workers' share of wages on relative wages, so identification resembles reverse regressions. We failed to find any explicit justification of this identification approach in the specific context of skill substitution. The studies in this group often use relatively disaggregated data, so they might implicitly rely on the same assumption as the first group of studies described above. In addition, estimates from both translog cost and production functions (given the flexibility of the translog function, often means or medians of many potential values) are less commonly accompanied by standard errors or other statistics from which standard errors can be computed, which effectively prevents us from attempting any analysis of publication bias for this subgroup.

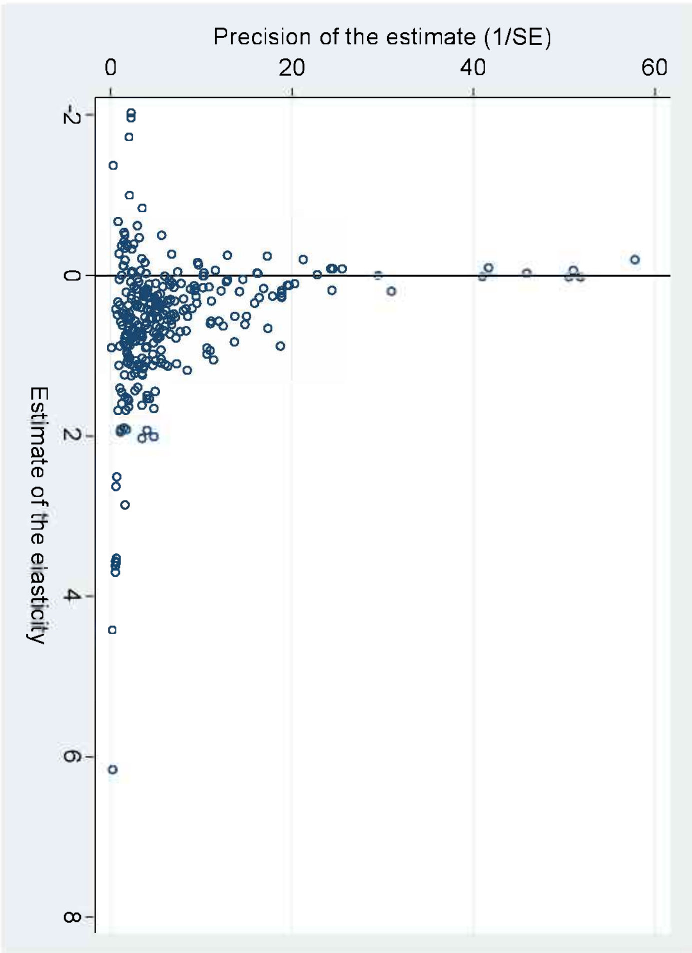
3.C Additional Material: Publication Bias

Here we present results for the sample of direct estimates, additional bias-correction techniques (for context, they are reported and discussed together with the techniques introduced in the main text), results for various subsets of the inverse estimates, and tests related to the assumption of conditional independence of estimates and standard errors. The additional techniques that we use are OLS meta-regression, the weighted average of adequately powered estimates (WAAP) by Ioannidis *et al.* (2017), and the stem-based technique by Furukawa (2020). WAAP first roughly estimates the underlying effect, then computes retrospective power for each reported estimate, drops estimates with less than 80% power, and computes the average of the remaining estimates weighted by inverse variance. The question is what rough estimate to choose in the first stage, and we follow the baseline specification of Ioannidis *et al.* (2017) by selecting the mean weighted by inverse variance. The stem-based technique builds on the funnel plot and exploits the trade-off between variance and bias: when only the most precise studies are included to compute the mean, publication bias is small, but it is inefficient to discard so many studies. Furukawa (2020) shows how to optimally balance bias and variance.

Figure 3.9 shows evidence of asymmetry in the funnel plot for direct estimates, a finding consistent with publication bias against negative and insignificant estimates of the elasticity. Table 3.10 shows the results of publication bias tests for direct and inverse elasticities. Regarding direct elasticities, all techniques find evidence of publication bias or, if they do not provide tests of the bias (WAAP and Stem), find corrected mean estimates much smaller than the uncorrected mean of 0.9. The use of study-level fixed or between effects does not change the conclusion. In the IV specification the instrument for the standard error is weak (the first-stage robust F-statistic is 6), so we report the two-step weak-instrument 95% confidence interval based on Andrews (2018). The overall message, consistent with the funnel plot, is that the corrected direct elasticity is zero. The only exception is the selection model by Andrews & Kasy (2019), which is more conservative in publication bias correction and suggests a mean of 0.4. But even this model finds that estimates significant at the 5% level are about twice as likely to be reported than insignificant estimates, and the implied exaggeration due to publication bias is more than twofold.

Regarding inverse elasticities, our results are similar. Publication bias is strong, and the corrected inverse elasticity is close to zero with the exception of the Andrews & Kasy (2019) model: even here, however, the implied elasticity of substitution exceeds 3. Note that the results for direct and inverse estimates are mutually inconsistent because they imply, respectively, zero and infinite elasticity of substitution. The identification assumptions in one of the streams of the literature are thus likely

Figure 3.9: The funnel plot suggests publication bias among direct estimates



Notes: When there is no publication bias, the funnel plots should be symmetrical. Outliers are excluded from the figure for ease of exposition but included in all statistical tests.

violated; as we explain in Section 3.B, we believe this is the case for most of the direct estimates. Another potential explanation is that both streams of the literature are identified reasonably well but attenuation bias drives the estimated coefficients to zero. While the sample of direct estimates is too limited to allow a meaningful analysis of attenuation bias, we find evidence consistent with the bias in the sample of inverse elasticities.

For the remaining analysis we only consider inverse estimates. Table 3.11 shows the results for the subsamples of OLS, IV, and natural experiments. The addition of OLS, WAAP and Stem techniques does not change the conclusions described in the main text (only the Stem method never finds a statistically significant corrected inverse elasticity). Unfortunately we only have 40 estimates from 6 natural experiments, so the power of the tests is low, but all techniques suggest strong publication bias and negligible corrected effects. Natural experiments are thus consistent with no causal effect of relative skill supply on the skill premium and therefore with infinite elasticity of substitution. We obtain similar results for OLS estimates—with the exception of the Andrews & Kasy (2019) model, which is once again less aggressive in correcting for publication bias. IV estimates of the negative inverse elasticity are different: they show less publication bias and more negative inverse elasticities, implying an elasticity around 4. The results are consistent with attenuation bias in the literature (IV estimates of inverse elasticities are larger in magnitude than OLS estimates) and little additional endogeneity bias (OLS estimates are similar to estimates from natural experiments).

Table 3.10: Tests point to strong publication bias, small corrected coefficients

Part 1: Direct elasticity				
Panel A: linear	OLS	FE	BE	IV
Publication bias	2.323***	1.574*	2.929***	1.534***
(*Standard error*)	(0.422)	(0.833)	(0.761)	(0.427)
	[-0.812, 5.227]			[0.919, 3.591]
				{0.459, 7.568}
Effect beyond bias	-0.0191	0.0906	0.0197	0.0963
(*Constant*)	(0.0843)	(0.122)	(0.0735)	(0.113)
	[-1.289, 0.524]			[-0.655, 0.341]
First-stage robust F-stat				6.05
Observations	311	311	311	309
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			2.333***	P=0.561
			(0.212)	(0.186)
Effect beyond bias	-0.0414	-0.0581	-0.0242	0.403**
	(0.0312)	(0.144)	(0.0195)	(0.205)
Observations	311	311	311	311
Part 2: Negative inverse elasticity				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-3.751***	-4.530***	-2.736***	-4.491***
(*Standard error*)	(0.735)	(1.311)	(0.781)	(1.161)
	[-5.54, -2.106]			[-6.985, -1.984]
				{-9.229, -2.127}
Effect beyond bias	-0.0822	-0.0259	-0.162***	-0.0491
(*Constant*)	(0.0781)	(0.0948)	(0.0470)	(0.103)
	[-0.348, 0.101]			[-0.381, 0.141]
First-stage robust F-stat				14.28
Observations	654	654	654	505
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-3.796***	P=0.385
			(0.296)	(0.110)
Effect beyond bias	-0.161***	0.102	-0.0768***	-0.300***
	(0.0220)	(0.208)	(0.0135)	(0.083)
Observations	654	654	654	654

Notes: Specifications in Panel A regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). WAAP = weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the method by Furukawa (2020), Endog. kink = the method by Bom & Rachinger (2019), Selection model = the method by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.11: IV estimation of the inverse elasticity shows less bias and a larger corrected effect

Part 1: OLS estimates of the inverse elasticity				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-5.393***	-5.804***	-4.277***	-6.962***
(*Standard error*)	(0.856)	(1.999)	(1.266)	(1.694)
	[-7.072, -3.165]			[-11.770, -2.494]
				{-11.972, -3.133}
Effect beyond bias	-0.0420	-0.0207	-0.0965	0.0103
(*Constant*)	(0.0757)	(0.103)	(0.0627)	(0.104)
	[-0.339, 0.134]			[-0.331, 0.214]
First-stage robust F-stat				46.17
Observations	347	347	347	251
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-5.465***	P=0.468
			(0.540)	(0.139)
Effect beyond bias	-0.144**	0.102	-0.0361**	-0.289**
	(0.0252)	(0.201)	(0.0191)	(0.113)
Observations	347	347	347	347
Part 2: IV estimates of the inverse elasticity				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-1.489***	-2.287**	-0.923	-0.553
(*Standard error*)	(0.577)	(0.843)	(1.365)	(0.681)
	[-2.962, 0.913]			[-1.913, 1.078]
				{-1.991, 0.748}
Effect beyond bias	-0.252**	-0.149	-0.297**	-0.400***
(*Constant*)	(0.123)	(0.109)	(0.115)	(0.114)
	[-0.561, 0.046]			[-0.719, 0.175]
First-stage robust F-stat				69.98
Observations	264	264	264	212
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-1.485***	P=0.336
			(0.268)	(0.093)
Effect beyond bias	-0.330***	-0.151	-0.252***	-0.333***
	(0.0308)	(0.195)	(0.0246)	(0.058)
Observations	264	264	264	264

Continued on next page

Table 3.11: IV estimation of the inverse elasticity shows less bias and a larger corrected effect (continued)

Part 3: Natural experiments				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-3.282***	-3.557***	-1.874*	-3.176***
(*Standard error*)	(0.880)	(0.0178)	(0.682)	(0.853)
	[-4.801, -1.610]			[-4.854, -1.407]
				{-4.653, -1.444}
Effect beyond bias	0.0116	0.0496***	-0.121	-0.00307
(*Constant*)	(0.0322)	(0.00246)	(0.0824)	(0.0297)
	[-0.672, 0.810]			[NA, NA]
First-stage robust F-stat				260.41
Observations	40	40	40	40
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-3.115***	P=0.187
			(0.343)	(0.075)
Effect beyond bias	-0.0146	-0.162	0.00302	-0.009
	(NA)	(0.162)	(0.0280)	(0.066)
Observations	40	40	40	40

Notes: Specifications in Panel A regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). WAAP = weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the method by Furukawa (2020), Endog. kink = the method by Bom & Rachinger (2019), Selection model = the method by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 3.12, Table 3.13, and Table 3.14 we examine other subsamples: developed countries vs. developing countries, elasticities estimated at the country level vs. at the regional level, one-level CES functions vs. a multilevel CES functions. Once again the addition of OLS, WAAP and Stem does not change our conclusions. The results suggest that elasticities tend to be larger for developed countries (above 4) than developing countries (around 2.5), and publication bias is stronger for the former group, which displays a corrected inverse elasticity closer to zero. Next, our results suggest that elasticities estimated at the country level are smaller than those estimated at the regional level, but we only have 93 estimates for the latter group. Finally, both one-level and multilevel CES functions seem to yield similar elasticities.

Table 3.12: Publication bias tests for subsamples of inverse elasticities estimated for developed and developing countries

Part 1: Inverse elasticities estimated for developed countries				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-3.744***	-4.526***	-2.748***	-5.391***
(*Standard error*)	(0.703)	(1.320)	(0.634)	(1.357)
	[-5.371, -2.073]			[-8.829, -2.605]
				{-14.237, -2.345}
Effect beyond bias	-0.0323	0.0194	-0.122***	0.0674
(*Constant*)	(0.0744)	(0.0875)	(0.0372)	(0.0675)
	[-0.323, 0.151]			[-0.244, 0.203]
First-stage robust F-stat				8.72
Observations	418	418	418	299
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-3.771***	P=0.433
			(0.349)	(0.109)
Effect beyond bias	-0.0838***	0.102	-0.0276*	-0.258***
	(0.0285)	(0.171)	(0.0141)	(0.088)
Observations	418	418	418	418
Part 2: Inverse elasticities estimated for developing countries				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-0.456	-1.204	0.581	0.193
(*Standard error*)	(0.690)	(0.803)	(1.913)	(1.023)
	[-1.679, 2.875]			[-2.334, 5.368]
				{-1.743, 2.102}
Effect beyond bias	-0.404***	-0.349***	-0.478***	-0.457***
(*Constant*)	(0.0803)	(0.0593)	(0.149)	(0.0546)
	[-0.804, -0.0323]			[-0.837, -0.397]
First-stage robust F-stat				275.96
Observations	151	151	151	128
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-0.453	P=0.654
			(0.446)	(0.442)
Effect beyond bias	-0.423***	-0.391***	-0.404***	-0.425***
	(0.0188)	(0.0731)	(0.0261)	(0.030)
Observations	151	151	151	151

Notes: Specifications in Panel A regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). WAAP = weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the method by Furukawa (2020), Endog. kink = the method by Bom & Rachinger (2019), Selection model = the method by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.13: Publication bias tests for subsamples of inverse elasticities estimated at the country and regional level

Part 1: Inverse elasticities estimated at the country level				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-2.861***	-2.950***	-2.830***	-2.870**
(*Standard error*)	(0.689)	(0.907)	(0.878)	(1.207)
	[-4.799, -1.05]			[-5.694, 0.229]
				{-3.971, -1.146}
Effect beyond bias	-0.193**	-0.187***	-0.188***	-0.238***
(*Constant*)	(0.0753)	(0.0674)	(0.0531)	(0.0916)
	[-0.437, -0.045]			[-0.448, -0.063]
First-stage robust F-stat				7.87
Observations	555	555	555	406
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-2.828***	P=0.436
			(0.302)	(0.101)
Effect beyond bias	-0.281***	-0.0884	-0.194***	-0.311***
	(0.0175)	(0.0923)	(0.0151)	(0.096)
Observations	555	555	555	555
Part 2: Inverse elasticities estimated at the regional level				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-2.442***	-2.153***	-2.098***	-2.482***
(*Standard error*)	(0.304)	(0.480)	(0.400)	(0.280)
	[-3.989, -0.991]			[-3.640, -1.121]
				{-3.213, -1.662}
Effect beyond bias	-0.0301	-0.0563	-0.0299	-0.0265***
(*Constant*)	(0.0186)	(0.0434)	(0.0193)	(0.00791)
	[-0.121, 0.155]			[-0.055, -0.004]
First-stage robust F-stat				52.11
Observations	93	93	93	93
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-2.440***	P=0.105
			(0.198)	(0.059)
Effect beyond bias	-0.0618**	-0.0668	-0.0304***	-0.162
	(0.0233)	(0.0793)	(0.0102)	(0.114)
Observations	93	93	93	93

Notes: Specifications in Panel A regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). WAAP = weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the method by Furukawa (2020), Endog. kink = the method by Bom & Rachinger (2019), Selection model = the method by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.14: Publication bias tests for subsamples of inverse elasticities estimated using one-level and multilevel CES functions

Part 1: Inverse elasticities estimated using one-level CES funtions				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-4.815***	-5.376***	-2.066	-6.121***
(*Standard error*)	(0.806)	(1.156)	(1.318)	(1.697)
	[-6.282, -1.797]			[-14.650, -2.333]
				{-9.133, -2.670}
Effect beyond bias	0.0695	0.114	-0.199*	0.197***
(*Constant*)	(0.126)	(0.0923)	(0.103)	(0.0234)
	[-0.739, 2.054]			[-3.071, 1.167]
First-stage robust F-stat				6,322.62
Observations	198	198	198	149
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-4.903***	P=0.507
			(0.549)	(0.191)
Effect beyond bias	0.144***	0.130	0.0778***	-0.371***
	(0.0209)	(0.152)	(0.0256)	(0.118)
Observations	198	198	198	198
Part 2: Inverse elasticities estimated using multilevel CES funtions				
Panel A: linear	OLS	FE	BE	IV
Publication bias	-3.289***	-3.370***	-3.331***	-3.345***
(*Standard error*)	(0.685)	(0.744)	(1.086)	(1.031)
	[-5.355, -1.709]			[-5.774, -.3949]
				{-4.595, -1.942}
Effect beyond bias	-0.144**	-0.139***	-0.145**	-0.177**
(*Constant*)	(0.0591)	(0.0513)	(0.0559)	(0.0863)
	[-0.372, -0.034]			[-0.465, -0.021]
First-stage robust F-stat				11.49
Observations	444	444	444	348
Panel B: nonlinear	WAAP	Stem method	Endog. kink	Selection model
Publication bias			-3.274***	P=0.350
			(0.328)	(0.090)
Effect beyond bias	-0.161***	-0.0860	-0.144***	-0.261**
	(0.0220)	(0.174)	(0.0147)	(0.104)
Observations	444	444	444	444

Notes: Specifications in Panel A regress estimates on standard errors (weighted by inverse variance). Standard errors, clustered at the study level, are in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. FE = study fixed effects. BE = study between effects. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. In curly brackets we show the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018). WAAP = weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the method by Furukawa (2020), Endog. kink = the method by Bom & Rachinger (2019), Selection model = the method by Andrews & Kasy (2019), P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.15: Regressing estimates on standard errors when p-value < 0.005

	All inverse	OLS method	IV method	Natural experiment	Developed country
Standard error	-4.864***	-6.208***	-0.947	-5.379***	-4.903***
	(0.941)	(1.227)	(0.922)	(0.368)	(0.816)
	[-7.376, -2.372]	[-9.257, -3.210]	[-2.848, 3.234]	[-6.323, -4.835]	[-6.902, -3.053]
Observations	368	237	115	15	222
	Developing country	Country estimate	Region estimate	One-level CES function	Multilevel CES function
Standard error	-0.848	-4.907***	-3.433***	-1.534	-5.339***
	(1.013)	(1.143)	(0.188)	(1.802)	(1.025)
	[-2.295, 4.223]	[-8.007, -2.069]	[-4.908, -1.406]	[-8.816, 2.251]	[-8.110, -2.947]
Observations	94	333	35	101	264

Notes: The response variable is the estimate of the negative inverse elasticity. The constant is included in the regressions but not reported in the table. Standard errors, clustered at the study level, are shown in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.* 2018) are in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.16: Specification test for the Andrews & Kasy (2019) model

	All inverse	OLS method	IV method	Natural experiment	Developed country
Correlation	0.515	0.265	0.635	0.947	0.558
	[0.436, 0.605]	[0.160, 0.392]	[0.499, 0.743]	[0.928, 0.976]	[0.462, 0.667]
Observations	654	347	264	40	418
	Developing country	Country estimate	Region estimate	One-level CES function	Multilevel CES function
Correlation	0.330	0.856	0.476	0.199	0.592
	[0.159, 0.482]	[0.775, 0.95]	[0.384, 0.568]	[0.047, 0.342]	[0.498, 0.709]
Observations	151	555	93	198	444

Notes: Following Kranz & Putz (2022), the table shows, for various subsets of the literature, 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 & Kasy (2019) model. If the assumptions of the model hold, the correlation is zero. Bootstrapped 95% confidence interval in parentheses.

3.D Additional Material: Heterogeneity

The studies estimating the elasticity differ in so many dimensions that it is unfeasible to control for all potential differences. In the publication bias section we used study-level fixed effects, which capture study idiosyncrasies but not the characteristics of individual estimation specifications. At the risk that we still omit some characteristics others would find relevant—the list of potential ones is unlimited—, we identify 28 main characteristics (and consequently, to avoid the dummy trap, codify 24 explanatory variables to be used in model averaging) which we distribute for ease of exposition into five categories: data characteristics, structural variation, design of the production function, estimation technique, and publication characteristics. Table 3.17 lists all the codified characteristics, provides their definitions, and gives summary statistics including the simple mean, standard deviation, and mean weighted by the inverse of the number of estimates reported in a study. Given the number of estimates that we collect, the construction of the dataset required manual collection of about 30,000 data points by three of the co-authors upon carefully reading the primary studies.

Table 3.17: Description and summary statistics of regression variables

Variable	Description	Mean	SD	WM
Inverse elasticity	Estimate of the negative of the inverse elasticity of substitution between the skilled and unskilled labor (response variable).	-0.543	0.436	-0.515
Standard error (SE)	Standard error of the estimated inverse elasticity. The variable is important for gauging publication bias.	0.190	0.240	0.184
Data characteristics				
Higher frequency	=1 if higher than annual frequency of the data is used; typically monthly, quarterly, or semi-annual.	0.083	0.275	0.134
Annual frequency	=1 if annual frequency of the data is used in the estimation (reference category for data frequency).	0.784	0.412	0.796

Continued on next page

Table 3.17: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD	WM
Lower frequency	=1 if lower than annual frequency of the data is used; typically three, five, or ten years.	0.133	0.340	0.070
Micro data	=1 if micro-level data (unit = single worker or firm) are used in the estimation.	0.257	0.437	0.267
Sectoral data	=1 if sector-level data (unit = sector) are used in the estimation.	0.164	0.370	0.122
Aggregated data	=1 if aggregated data (unit = economy or regions) are used in the estimation (reference category for the type of data aggregation).	0.580	0.494	0.611
Cross-section	=1 if cross-sectional data are used; =0 if time-series or panel data are used.	0.076	0.266	0.138
Structural variation				
United States	=1 if the country for which the elasticity is estimated is the United States.	0.410	0.492	0.422
Developing country	=1 if a developing country is considered.	0.231	0.422	0.177
Manufacturing sector	=1 if the elasticity is estimated for the manufacturing sector, =0 if another sector is considered.	0.145	0.353	0.058
Design of production function				
One-level CES function	=1 if a one-level CES form of the production function is used in the estimation (reference category for the functional form).	0.321	0.467	0.378
Multilevel CES function	=1 if a multilevel CES form of the production function is used in the estimation.	0.679	0.467	0.622
Time control	=1 if time control is included in the model (capturing, e.g., technological change).	0.544	0.498	0.640
Location control	=1 if location/unit control is included (capturing spatial variation).	0.142	0.350	0.149
Macro control	=1 if macroeconomic indicators are included.	0.086	0.280	0.084
Age control	=1 if a control for the age of workers is included.	0.098	0.297	0.179
Capital control	=1 if a capital-related control is included (capturing changes in capital stock under a capital-skill complementarity technology).	0.214	0.410	0.148
Estimation technique				

Continued on next page

Table 3.17: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD	WM
Dynamic model	=1 if the model form used for estimation is dynamic (VAR, ECM, VECM, PAD, ADL, DLTM, DOLS).	0.076	0.266	0.168
Unit fixed effects	=1 if the model is estimated in first differences or cross-sectional fixed effects are considered.	0.517	0.500	0.430
Time fixed effects	=1 if time fixed effects are included in the model.	0.240	0.427	0.165
OLS method	=1 if the ordinary least squares method or its variations (LS, DOLS, WLS, GLS) are used for estimation (reference category for the method variables).	0.531	0.499	0.611
IV method	=1 if instrumental variables are used, including 2SLS, 3SLS, and GMM.	0.404	0.491	0.270
Natural experiment	=1 if the study uses data from a natural experiment (e.g., has arguably exogenous variation in the relative supply of skilled labor).	0.061	0.240	0.089
Publication characteristics				
Impact factor	The discounted recursive RePEc impact factor of the outlet.	0.846	1.169	1.385
Citations	The logarithm of the number of per-year citations of the study in Google Scholar.	1.612	1.434	2.126

Notes: Table only includes estimates of the inverse elasticity for which standard errors are reported. SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study, CES = constant elasticity of substitution, VAR = vector autoregression, ECM = error correction model, VECM = vector error correction model, PAD = partial adjustment model, ADL = autoregressive distributed lag model, DLTM = distributed lag and trend model, DOLS = dynamic ordinary least squares, WLS = weighted least squares, GLS = generalized least squares, 2SLS = two-stage least squares, 3SLS = three-stage least squares, GMM = generalized method of moments.

Data characteristics: The studies in our sample differ in the type of data used to produce estimates of the elasticity. An important aspect is data frequency. With higher frequencies, transitory variation is often present and, if not accounted for, it can generate a biased estimate of the long-run elasticity of substitution (Chirinko & Mallick 2017). Four fifths of the estimates in our sample employ *annual* data; higher frequencies such as monthly, quarterly, or semi-annual appear relatively scarcely.

Another challenge that the researchers have to face is that of data aggregation. Hamermesh (1996) classifies empirical studies into three main groups based on the level of data aggregation. First, there are studies using *aggregated data*, where the unit of observation is the economy or region. Second, aggregation can be conducted at the level of industries (captured by the variable *Sectoral data*). The third group consists of studies where firms or individuals are used as units of observation (*Micro data*).

There are several potential problems with data aggregation. For instance, Hamermesh (1996) criticizes the use of linear aggregation techniques for aggregating nonlinear relationships. Even with the assumption of identical technologies in all firms, one cannot be sure that parameters in the estimated equations are the same for the particular firm and for the aggregated case. Moreover, aggregating workers into groups means implicitly assuming that these workers are very close p-substitutes or q-complements. Furthermore, Broadstock *et al.* (2007) warn that the estimated elasticity involving an aggregate is not necessarily a weighted average of the elasticities for the disaggregated inputs. A practical issue with aggregated data is that fewer observations used in regressions are usually linked with lower precision and that measurement error can differ from that in disaggregated data, which has consequences for both publication and attenuation bias. Another important aspect of data is their dimension: if purely *cross-sectional data* are used or if the time dimension is also taken into account. Hamermesh (1996, p. 63) notes that "there is nothing inherently more attractive in cross-sections or time-series data. Rather, the choice depends on the degree of spatial aggregation in each type of available data." In practice, time series at the micro level are rare, and cross-sectional data generally enable greater disaggregation.

Structural variation: Inherent differences in the elasticity among countries and sectors could give rise to another source of heterogeneity. A large part of our sample consists of elasticities computed for the United States (about 40%). The strong consensus in the literature about the elasticity lying between 1 and 2 is to a large extent derived from the US studies by Katz & Murphy (1992), Ciccone & Peri (2005), Autor *et al.* (2008), and Goldin & Katz (2009). Evidence on structural variation has been rather rare in the literature. Psacharopoulos & Hinchliffe (1972) report larger estimates for developed countries compared to developing ones, while Tinbergen (1974) finds the values between 0.4 and 2 for both developing and developed countries. Later studies on *Developing countries*, such as Behar (2010) or Manacorda *et al.* (2010), suggest values between 2 and 4. On balance, according to our reading of the literature the prevailing view is still that the elasticity is larger in more developed countries (Foldvari & van Leeuwen 2006).

Some authors, such as Blankenau & Cassou (2011), suggest that manufacturing and skilled services (financial or health services, for example) often stand out in the industry-specific analyses. These sectors, with a shifting demand to skilled labor and heavier on specific skill-sets (Berman *et al.* 1994), may display structurally different elasticities. We create a separate dummy for manufacturing, for which we have enough observations.

Design of the production function: Researchers typically assume *one-level CES* (constant elasticity of substitution) production function. But other functional forms are used as well, including most prominently the multilevel (or nested) CES function. For the sake of simplicity, some authors consider solely equation (3.2), which treats skilled and unskilled labor as the only factors of production. In this form the elasticity is constant irrespective of changes in relative labor supply (Ciccone & Peri 2005) and can be derived from the parameter ρ as $\sigma = 1/(1-\rho)$. Under the CES framework, more production factors can be nested (*Multilevel CES*) and there are many ways to do so.

Most often, three production factors are considered in estimation: skilled labor, unskilled labor, and capital. One stream of the literature assumes production to be a CES function of capital and labor at the first level and further decomposes labor into skilled and unskilled at the second level via the Cobb-Douglas specification (therefore, the elasticity of substitution between capital and labor is restricted to one; Avalos &

Savvides 2006). Another stream of the literature assumes a CES function with capital and labor at the first level and further decomposes labor into skilled and unskilled parts at the second level via another CES specification (as in Borjas 2003; Borjas & Katz 2007). Finally, some studies apply alternative nesting schemes with a CES function of capital, skilled labor, and unskilled labor at the first level; at the second level, skilled workers are divided into more groups according to their specific skills via a CES specification (Manacorda *et al.* 2010). Among the more complex nesting structures is the one used by Krusell *et al.* (2000) and followed by Lindquist (2005) and Dupuy (2007). They employ four production factors (capital structure, capital equipment, skilled labor, and unskilled labor) and a three-level nesting structure.

Multiple control variables are commonly employed in the basic specification of the production functions described above. These variables capture different characteristics of either workers or labor markets. The most frequent one is *time control* capturing potential technological changes that affect the demand for skills; these controls are used in about half of the regressions in our sample. Other variables control for the location (Acemoglu 2002), different macroeconomic circumstances such as the level of minimum wage, unemployment rate, and labor market reforms (Manacorda *et al.* 2010; Autor *et al.* 2008), and socioeconomic factors such as city size, college share, and union membership (Freeman & Medoff 1982; Card 2009). The authors of primary studies also capture industry differences, variations in age cohorts, and capital stock.

Estimation techniques: We control for models that are *dynamic*, thus account for the fact that the elasticity may change in response to shocks (estimated in models such as vector autoregression, partial adjustment model, or distributed lag model, among others). We codify studies that account for *unit fixed effects* either using unit dummies or first differences. This method controls for persistent features that could affect the level of skill (or the level of technology used in firms) in specific cohorts of labor force: features such as location (Borjas & Katz 2007), degree (Kawaguchi & Mori 2016), age (Angrist 1995), and industry (Razzak & Timmins 2008). On the other hand, we also codify studies that account for *time fixed effects*. This method controls for temporal dynamics of unobservable factors changing in time that could affect the skill-biased technical change.

A notorious issue in the empirical literature estimating substitution elasticities is that of potential endogeneity bias. Researchers try to address this problem by instrumenting labor supply, but good instruments are hard to come by. An example of a suitable *instrument* can be found in Ciccone & Peri (2005), who use state and year specific compulsory school attendance and child labor laws as instruments for relative labor supply of more educated workers. Such an approach also corrects for potential attenuation bias resulting from measurement error. We also control for studies using *natural experiments*, which we define as studies having access to arguably exogenous variation in the relative supply of skill. The most prominent recent example is Carneiro *et al.* (2022), who exploit the construction of new colleges in Norway in the 1970s. Natural experiments tackle endogeneity, but generally with the exception of attenuation bias. In addition, natural experiments typically cover only a short period of time, and their estimates might be biased to zero due to adjustment lags because they might not be able to fully capture the long-run effect of labor supply on prices.

Publication characteristics: To account for aspects of quality not captured by the variables introduced above, we employ two additional variables. First, we use the number of citations taken from Google Scholar (variable *Citations*) normalized by the number of years since the first draft of the study appeared in Google Scholar. Second, we use the RePEc recursive discounted impact factor, which is available for journal articles as well as working papers (variable *Impact factor*).

3.E Diagnostics and Robustness Checks of BMA

Table 3.18: Diagnostics of the benchmark BMA estimation (UIP and dilution priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
15.2681	$3 \cdot 10^5$	$1 \cdot 10^5$	1.79 mins	77,558
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	46.00%	100%	0.9975	654
Model prior	*g-prior*	*Shrinkage-stats*		
Uniform/12	UIP	Av = 0.9985		

Notes: We employ the combination of unit information prior recommended by (Eicher *et al.* 2011) and dilution prior suggested by George (2010), which accounts for collinearity.

Figure 3.10: Benchmark BMA model size and convergence (UIP and dilution priors)

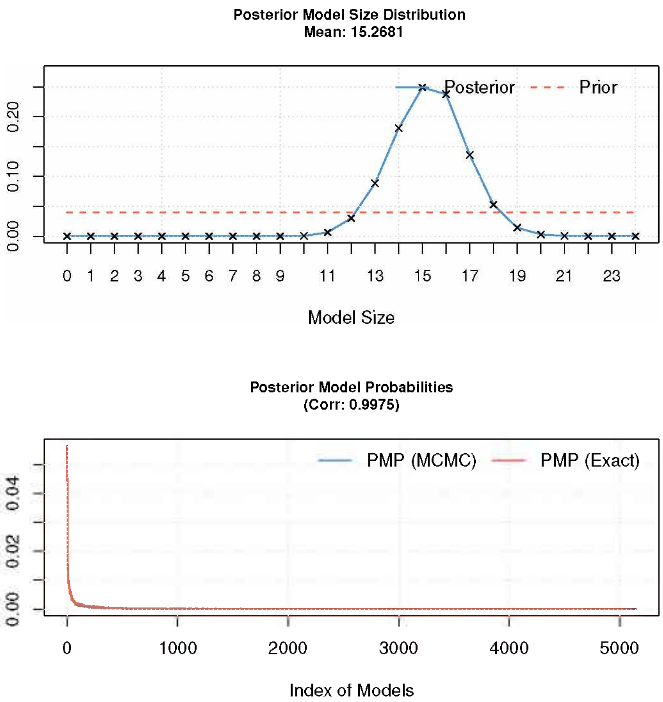
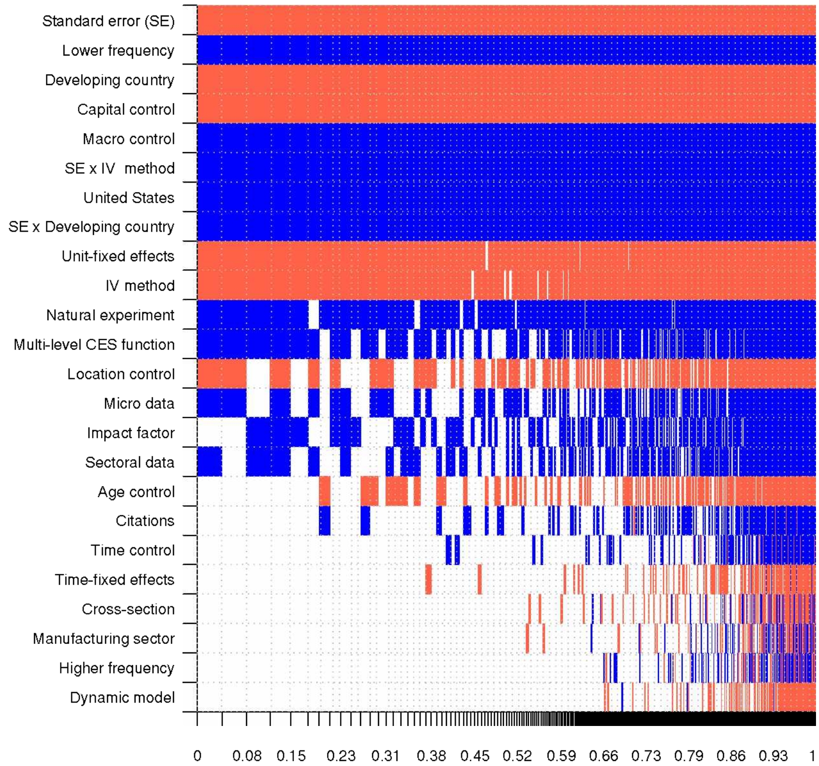


Table 3.19: Why elasticities vary (alternative priors)

Response variable: Inverse elasticity	Bayesian model averaging (uniform prior)			Bayesian model averaging (BRIC prior)			Bayesian model averaging (hyper-g prior)		
	P. mean	P. SD	PIP	P. mean	P. SD	PIP	P. mean	P. SD	PIP
Constant	-0.19	NA	1.00	-0.20	NA	1.00	-0.21	NA	1.00
Standard error (SE)	-3.62	0.89	1.00	-3.62	0.84	1.00	-3.64	0.80	1.00
SE * IV method	2.31	0.50	1.00	2.35	0.48	1.00	2.35	0.44	1.00
SE * Developing country	2.22	0.61	0.99	2.24	0.59	1.00	2.23	0.56	1.00
Data characteristics									
Higher frequency	0.00	0.01	0.05	0.00	0.02	0.08	-0.01	0.04	0.38
Lower frequency	0.25	0.04	1.00	0.26	0.04	1.00	0.27	0.04	1.00
Micro data	0.05	0.05	0.50	0.06	0.05	0.65	0.08	0.04	0.93
Sectoral data	0.05	0.06	0.46	0.07	0.06	0.61	0.10	0.05	0.91
Cross-section	0.00	0.01	0.06	0.00	0.01	0.10	-0.01	0.03	0.39
Structural variation									
United States	0.10	0.03	1.00	0.10	0.03	1.00	0.10	0.02	1.00
Developing country	-0.21	0.04	1.00	-0.21	0.04	1.00	-0.20	0.04	1.00
Manufacturing sector	0.00	0.02	0.05	0.00	0.02	0.09	-0.01	0.05	0.38
Design of production function									
Multilevel CES function	0.05	0.04	0.67	0.05	0.04	0.79	0.06	0.03	0.95
Time control	0.00	0.01	0.08	0.00	0.01	0.11	0.00	0.02	0.37
Location control	-0.08	0.09	0.53	-0.10	0.08	0.65	-0.13	0.07	0.91
Macro control	0.19	0.04	1.00	0.19	0.04	1.00	0.19	0.03	1.00
Age control	-0.02	0.03	0.31	-0.02	0.03	0.36	-0.02	0.03	0.60
Capital control	-0.39	0.03	1.00	-0.39	0.03	1.00	-0.38	0.03	1.00
Estimation technique									
Dynamic model	0.00	0.01	0.04	0.00	0.02	0.07	-0.01	0.03	0.35
Unit fixed effects	-0.08	0.03	0.97	-0.08	0.02	0.99	-0.08	0.02	1.00
Time fixed effects	0.00	0.01	0.08	0.00	0.01	0.13	-0.01	0.02	0.43
IV method	-0.11	0.05	0.93	-0.12	0.04	0.96	-0.13	0.04	1.00
Natural experiment	0.19	0.09	0.91	0.19	0.08	0.92	0.17	0.07	0.96
Publication characteristics									
Impact factor	0.01	0.01	0.49	0.01	0.01	0.55	0.01	0.01	0.76
Citations	0.00	0.01	0.20	0.00	0.01	0.20	0.00	0.01	0.40
Studies	68			68			68		
Observations	654			654			654		

Notes: P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. In the first specification from the left we employ Bayesian model averaging combining the uniform model prior and the unit information g-prior recommended by Eicher *et al.* (2011). The second specification BRIC and Random = a g-prior by Fernandez *et al.* (2001) for parameters with the beta-binomial model prior (Ley & Steel 2009) for model space; this ensures that each model size has equal prior probability. The third specification uses a random model prior advocated by Ley & Steel (2009) and the data-dependent hyper-g prior suggested by Feldkircher & Zeugner (2012). All variables are described in Table 2.4.

Figure 3.11: Model inclusion in BMA (UIP and uniform priors)



Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 3.19. All variables are described in Table 3.17.

Table 3.20: Diagnostics of the BMA estimation (UIP and uniform priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
14.3781	$3 \cdot 10^5$	$1 \cdot 10^5$	1.45 mins	71,233
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	42.00%	100%	0.9966	654
Model prior	*g-prior*	*Shrinkage-stats*		
Uniform/12	UIP	Av = 0.9985		

Notes: We employ the priors suggested by Eicher *et al.* (2011), who recommend using the uniform model prior (each model has the same prior probability) and the unit information prior (the prior provides the same amount of information as one observation from the data).

Figure 3.12: BMA model size and convergence (UIP and uniform priors)

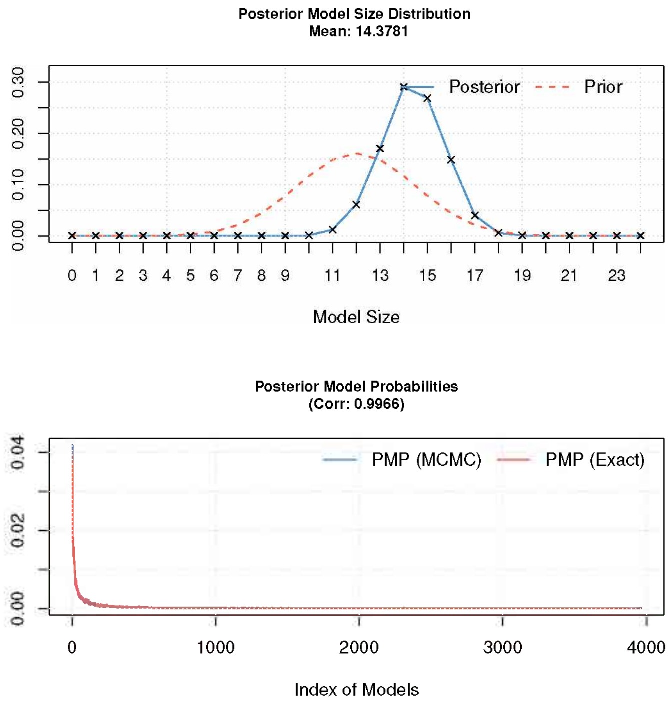
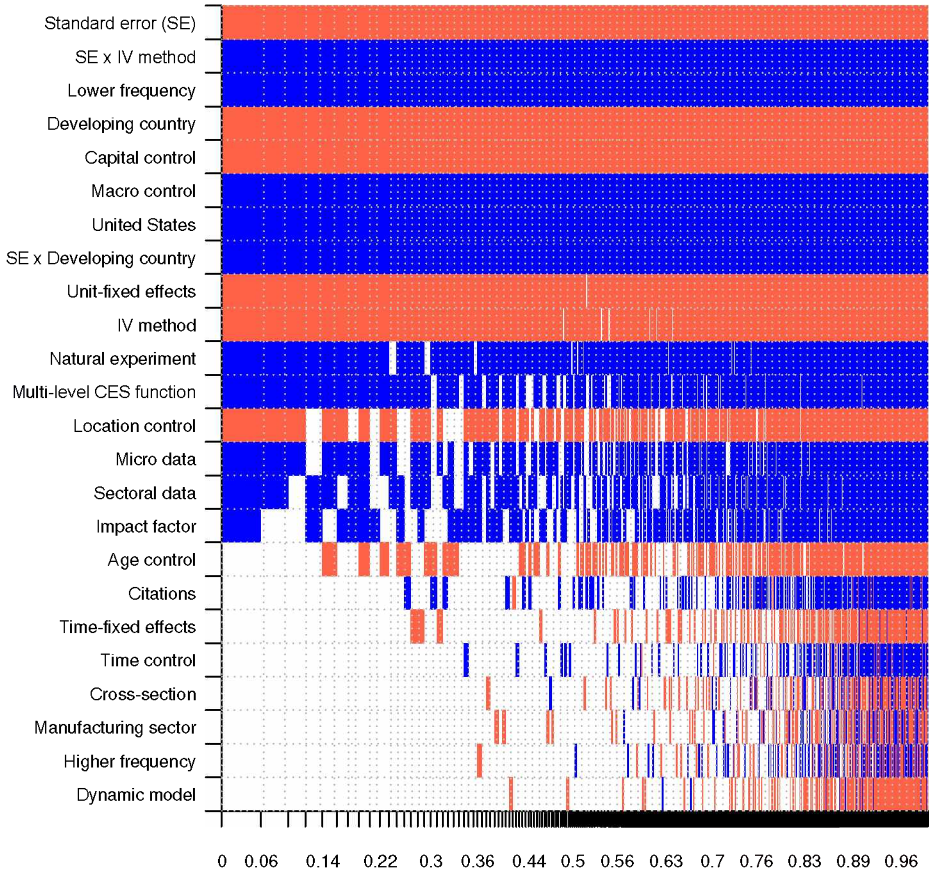


Figure 3.13: Model inclusion in BMA (BRIC and random priors)



Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 3.19. All variables are described in Table 2.4.

Table 3.21: Diagnostics of the BMA estimation (BRIC and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
15.2321	$3 \cdot 10^5$	$1 \cdot 10^5$	1.68 mins	77,762
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	46.00%	100%	0.9981	654
Model prior	*g-prior*	*Shrinkage-stats*		
Random/12	BRIC	Av = 0.9985		

Notes: The specification uses a BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure 3.14: BMA model size and convergence (BRIC and random priors)

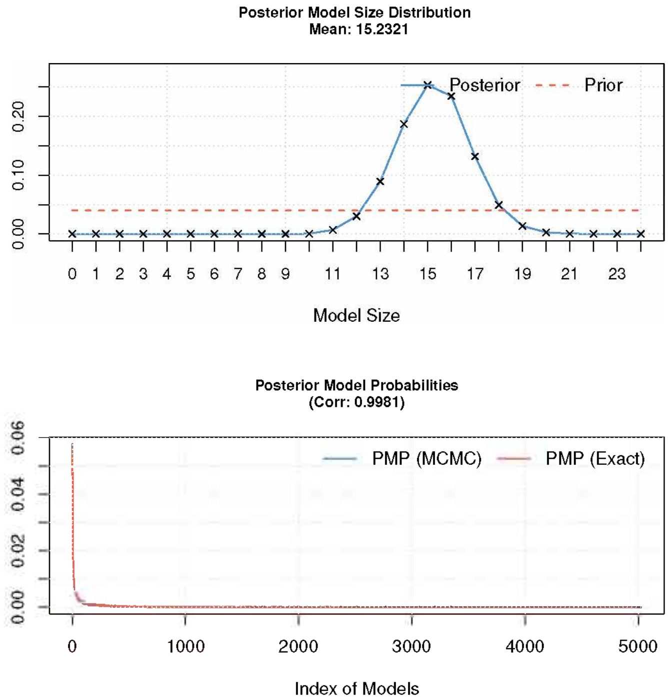
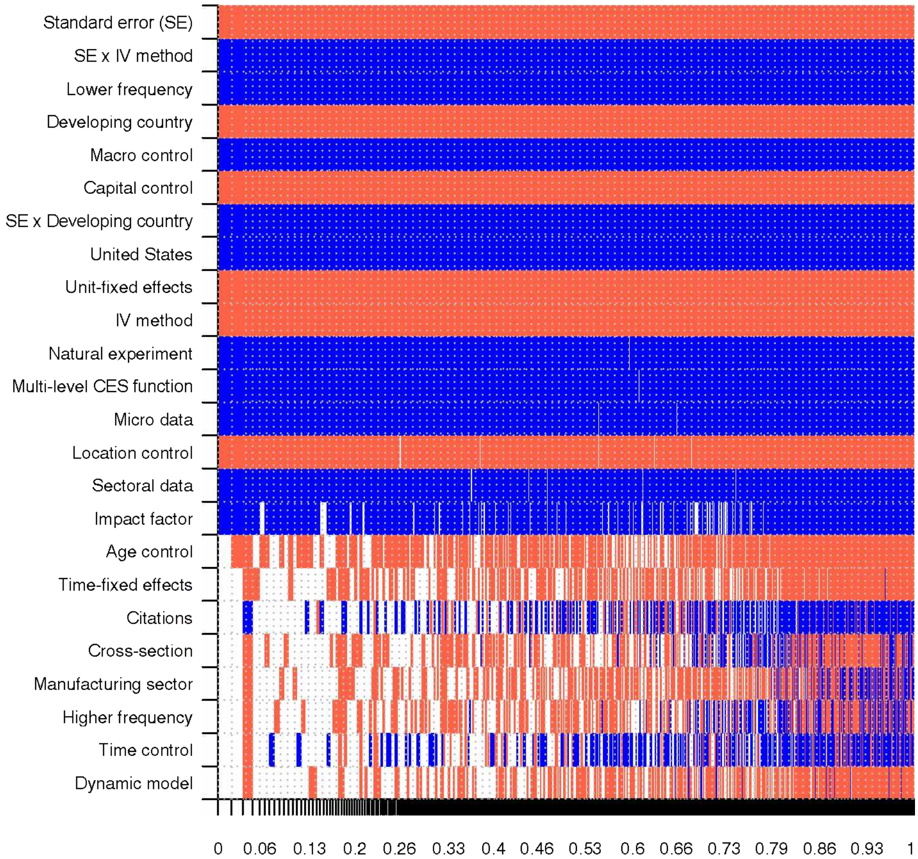


Figure 3.15: Model inclusion in BMA (hyper-g and random priors)



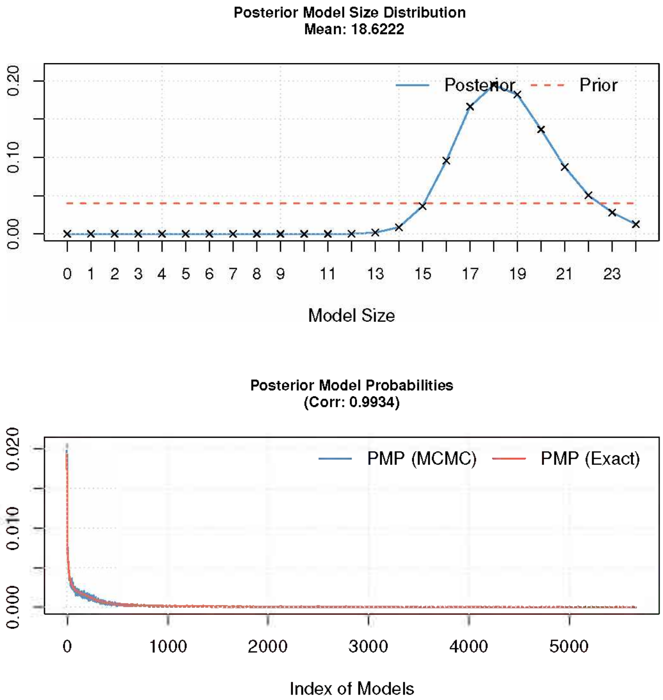
Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 3.19. All variables are described in Table 2.4.

Table 3.22: Diagnostics of the BMA estimation (hyper-g and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
18.6222	$3 \cdot 10^5$	$1 \cdot 10^5$	2.52 mins	113,680
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	68.00%	100%	0.9934	654
Model prior	*g-prior*	*Shrinkage-stats*		
Random/12	hyper (a=2.003058)	Av=0.9814, Stdev=0.0065		

Notes: The specification uses the data-dependent hyper-g prior suggested by Feldkircher & Zeugner (2012) and a random model prior advocated by Ley & Steel (2009).

Figure 3.16: BMA model size and convergence (hyper-g and random priors)



Chapter 4

Expected Returns on Higher Education in Russia after Unified State Exam Reform

Abstract

This paper investigates the impact of the Russian Unified State Exam (USE) reform on earnings for a sample of 13,790 individuals monitored via Russian Longitudinal Monitoring Survey (RLMS) between 1994 and 2020. Using a quasi-experimental design, I examine how the policy reform reallocated and enhanced human capital. The findings suggest that while the reform significantly affected all individuals graduating in major cities, smaller municipalities, and rural areas, it benefited only the treatment group in Moscow. Despite controlling for multiple factors, including flexible time trends, the analysis consistently shows a positive post-reform impact on earnings. The conservative estimate is approximately 8% for the treatment group and 5.5% for the control group.

Keywords: Human capital; unified state exam; return on education; schooling.

JEL Codes: I23, I26, J30

The third study is an unpublished solo manuscript. I thank Jan Janku, Nikolai Cook, Jerome Geyer-Klingeberg, Martin Gregor, Barbara Pertold-Gebicka, Milan Scasny, and Lenka Stastna for their valuable comments.

4.1 Introduction

The introduction of the Unified State Exam (USE) in 2009 completely transformed the admission process for all universities in Russia. This reform involves a single exam combining high school graduation and university admission content. It was expected that the USE would reallocate knowledge and skills better and induce better matching between the levels of students and universities (Ampilogov *et al.* 2013). The effect of USE reform has been widely debated and disputed both in the political arena and among the general public in Russia. Studies have assessed the effectiveness of specific educational strategies within the transition from high school to higher education in Russia and showed that this reform led to a threefold increase in geographic mobility rates among high school graduates from small cities and towns starting college in major cities (e.g., Francesconi *et al.* 2019). Prakhov & Yudkevich (2019) show that the choices of higher education and mobility have been influenced by family income and regional socioeconomic characteristics. Moreover, there is a lack of empirical research on the impact of the USE on the average salary in Russia due to an insufficient number of observational years. Given that the treatment group affected by the reform has been in the labor market since 2012, it is possible to assess whether these individuals fare better than do their peers. To my knowledge, this is the first study to assess the impact of the USE on average earnings using longitudinal data on 13,790 individuals from 1994 to 2020.

The decentralized admission exam for higher education that requires applicants to take university-specific entrance exams might impose certain aspects of incompatibility, such as additional application costs, time consumption, and limitation of choice. Although there is substantial friction within decentralized exams, the best higher education systems are found in countries that do not have standardized entrance tests, e.g., Japan, Finland, Israel and, until recently, South Korea and Brazil (e.g., Avery *et al.* 2014; Kerr *et al.* 2020). In the US, there is no centralized admission test for college; however, university candidates often have to complete college-specific requirements, such as admission essays and Scholastic Assessment Tests (SATs) (Goodman *et al.* 2020). Each university can find its preferred candidate according to its own expectations. Additionally, unified entrance tests might be considered a blanket way to test subject comprehension, as it induces students to concentrate on passing an examination; teachers might even teach students only a specific way in which to handle exams well, and thus, learning might be trivialized.

Although centralized state exams are the primary criterion for admission (such as in China, Taiwan, Italy, Belgium, Norway, and Germany), they are typically preferred by higher-ability students, whereas decentralized state exams are typically preferred by lower-ability students (Hafalir *et al.* 2018). Although the different outcomes between university-specific versus centralized university entrance exams have been investigated from a theoretical perspective (e.g., Chade *et al.* 2014; Che & Koh 2016), there is little empirical evidence of their returns on education. The adoption of USE reform in Russia provides an opportunity to evaluate the impact of the centralized exam on the effective matching of students with universities and on students' acquisition of skills and knowledge, ultimately leading to potential increases in their earnings and salaries.

The emerging consensus from previous studies is that university access was highly unequal prior to the 2009 reform. Specifically, students from lower socioeconomic backgrounds were significantly less likely to apply to college and earn a degree than were their peers with higher socioeconomic status. For example, an analysis of data from the 2006 Russian Longitudinal Monitoring Survey (RLMS) highlights that approximately 65% of 25 to 29 year olds who had a university degree reported that their father also possessed a university qualification, compared to a mere 20% among those whose fathers lacked such credentials (Francesconi *et al.* 2019). Additionally, prior to 2009, less than 20% of young Russians were born in the ten largest cities, including Moscow and St. Petersburg, yet this group constituted more than 60% of all university graduates. In contrast, individuals born in small cities, towns, and rural areas, which comprised approximately half of the population, accounted for merely one in ten graduates. This disproportionate representation of high school graduates from large cities among university students was linked to a sharp socioeconomic gradient.

To increase productivity in the labor market, countries must strike a balance between policies that promote flexibility, labor mobility, and job security. Higher education is crucial in enabling the younger generation to acquire new skills and adapt to the evolving demands of the labor market. By improving skill systems, distributing knowledge and expertise more equitably, and providing more equitable opportunities, the well-being of all individuals can be improved. In this study, I investigate the association between the allocation of skills and the absolute returns on higher education resulting from the USE. A previous study by Prakhov (2021) used longitudinal survey data from 1987 students across 46 regions in Russia to examine the determinants of expected returns on higher education, with the USE score, university, individual, and regional characteristics as predictors. My research builds upon this by analyzing longitudinal data from 1994 to 2020, encompassing 302,704 observations of 13,790 individuals across 32 states and seven federal districts. I take advantage from data on individual and family characteristics, as well as regional variation in economic and social factors and try to identify the factors responsible for the observed post-reform increase in earnings.

I find that, overall, the reform had a positive effect on human capital. My conservative estimates of the returns to schooling are 7% for the chosen treatment group and 5% for the chosen control group. My conclusions are based on the quasi-experimental design of the difference-in-differences technique as a baseline method and the propensity score matching as a robustness check to minimize the risk of selection bias. In addition, I exploit the individual and family characteristics that might have both direct and indirect effects on the expected returns on higher education. My findings indicate that, on average, females earn 30% less than males, and married individuals earn more than unmarried individuals. Furthermore, having a parent with a higher education diploma increases an individual's average salary by 6% for each parent, and if both parents possess a higher education diploma, the individual is expected to earn 13% more than their peers without such a family background. Additionally, having a child under six years old in the household has a more severe negative impact on average salary in major cities. The results remain robust across the estimation methodologies and alternative control groups.

The remainder of the paper is as follows. In section Section 4.2, I provide background information on the institutional context and details of the USE reform. The conceptual framework of how students make decisions about applying to university based on their abilities is outlined in section Section 4.3. In section Section 4.4, I use different methodological approaches to investigate the impact of the USE on earnings. Section Section 4.5 concludes. Additional information on the institutional context and supplementary results discussed throughout the paper can be found in the Section 4.A.

4.2 Institutional background

The transformations undergone by Russian universities in the wake of the collapse of the Soviet Union have been profound and far reaching. Moreover, the education system has undergone significant changes driven by a complex array of global and internal factors (Gounko & Smale 2007). These changes include the marketization and diversification of higher education institutions and the country's integration into the European higher education system through its participation in the Bologna Process. Additionally, the Russian government has made dramatic cuts in state financial support and introduced tuition fees, a significant departure from the previous Soviet-era policy of free education. These changes have had a significant impact on the structure and quality of higher education in Russia and on the opportunities available to students and the broader society.

The USE represents a single exam combining high school graduation exams and university entry tests. This exam was first introduced in 2001 as a pilot program in five states, and by 2009, it had become mandatory for all schools and universities across Russia. Prior to the USE introduction, each university conducted its own entrance examination, leading to concerns about the fairness and objectivity of the admission process. There were also issues surrounding the mismatch between the knowledge level provided at school and that demanded by universities, as well as surrounding the high costs associated with preparing for university entrance exams (Minina 2010). Additionally, there was significant inequality in access to higher education across different societal segments. With the goal of addressing these issues, the USE aimed to provide a more objective and fairer system of university admission while also encouraging a greater level of enrollment in universities. While the USE introduction has not been without controversy, it represents an important step forward in the ongoing evolution of the Russian education system.

Since the introduction of the USE in Russia, efforts have been made to improve the system through regular changes and updates. For instance, in 2015, two levels of mathematical tests were introduced to better match the needs of different students (RMES 2014). Additionally, significant changes were made to test materials in 2018, and minor adjustments and developments were implemented in 2019 (RMES 2018; 2019). However, despite these improvements, the higher education enrollment system is still undergoing stabilization, and thus, it is important to evaluate the actual impact of the USE on enrollment rates in Russia. This ongoing process provides an opportunity to identify any issues with the new system and make further improvements as necessary. Ultimately, the goal is to create a university admission process that is fair, objective, and accessible to all, regardless of socioeconomic status or background. With continued efforts and monitoring, the USE has the potential to play a key role in shaping the future of higher education in Russia.

According to Rosstat (2021), the total number of universities in Russia stood at 1,058 as of 2020, with the majority located in the Moscow federal district (107), St. Petersburg (48), and the Tatarstan Republic (27). Despite the large number of institutions, the government's expenditure on education as a share of gross domestic product (GDP) remains relatively low, at 4%. Every year, approximately 900,000 students are admitted to higher education institutions, with 65% of admissions being funded by the state. The total number of students enrolled in higher education in Russia currently stands at 4 million. In Moscow, the average annual tuition fee at universities is 278,000 Russian rubles (approximately 3,864 USD), making it the most expensive district for higher education in Russia, followed closely behind by

St. Petersburg, with an average tuition fee of 208,000 Russian rubles, while Tomsk comes in third place, with 164,000 Russian rubles (Rosstat 2021). These trends are also similar across other federal districts.

The leading fields of university study preferred by high school graduates in Russia in 2020 were engineering, technology, technical, and computer sciences, with males dominating these fields. Conversely, females make up a significantly higher proportion of students in fields such as economics and management, social sciences, and education and pedagogy. When selecting a university, students in Russia tend to prioritize education quality, institutional reputation, and professor expertise. Distance from home is also a significant factor for many students, with 21% considering proximity an important factor in their university selection process. The primary factors influencing students' choice of field are job prospects, opportunities for career growth, potential salary expectations, and the perception of the profession as respected and prestigious.

On one hand, the introduction of the USE has simplified the university entry process, reduced transaction costs for students, and allowed them to apply to multiple universities at the same time (Ampilogov *et al.* 2013). Additionally, USE scores proved to be a good predictor of talented applicants (Khavenson & Solovyova 2014). The USE has also made the admission process more transparent and significantly increased the mobility of students from small towns to big cities (e.g., Fedotova & Chigisheva 2010; Francesconi *et al.* 2019). On the other hand, the USE reforms have been criticized for being largely ineffective in some states (e.g., Luk'yanova 2012). Critics argue that the USE test is not able to reveal the level of knowledge and that the learning process is reduced to simply preparing for the exam rather than gaining deep knowledge of the subject. The research on the impact of unified exams on higher education is still in its early stages. Some studies, such as Ampilogov *et al.* (2013) or Francesconi *et al.* (2019), suggest that the USE in Russia has had a positive effect on students' mobility and increased the likelihood of applying to multiple universities. Despite the USE's significant role in higher education policies, there is still no consensus on its actual impact on student learning and mobility.

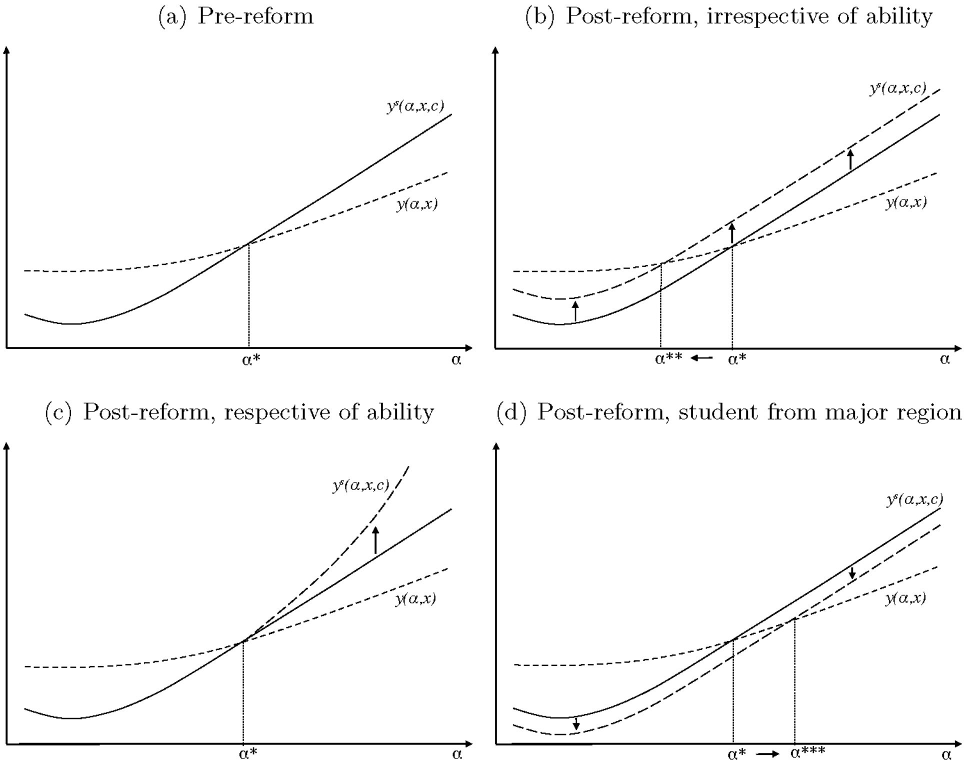
4.3 Research design

I develop a simple conceptual framework to understand how USE reform has influenced the average salary of those affected by it. My setup is based on the work by Francesconi *et al.* (2019), who they assess the impact of the USE on migration. I restructured the model to determine the effect of the USE on return on education. A primary focus of analysis is to examine how the USE reform has affected the education return and if it indeed reduced the cost of applications. I investigated whether the policy enabled high school students to better match with institutions that fit their interests and abilities, ultimately resulting in higher average salaries.

My analysis consists of several assumption. I assume a student has two choices: first, to attend a university in their city or the nearest city, and second, to relocate to a major city, particularly one with elite institutions. The student aims to maximize their future earnings ($y(a,x)$), which are determined by their ability, a, as well as their individual, family, and regional characteristics, x.

To specify the pre-reform conditions of a student, I assume that there is no unified state exam, and each university has its own entrance exams. If the student decides to move to another major city for the elite institution, there are additional costs related to extra effort to prepare for each college, additional tutoring, time, and traveling to the major cities for the entrance exam. Moreover, if they are admitted, they have an extra cost to move and start living in a major city. All these costs are denoted as c, the total cost. Since the student cannot attend all entrance exams at once, they have limited choices. To make a decision on whether to take the entrance exam for an elite university, the student compares the benefits and costs of both options. If they decide to stay in their own region, they are not exposed to additional costs, and their earning function is denoted as $y(a,x)$. However, if they decide to take an entrance exam for an elite university and move to a major city, the earning function is given by $y^s(a,x,c)$, which increases in a and x but decreases in c. Additionally, if student's ability is higher, the associated cost is reduced, which means they spent less effort preparing. Therefore, $y^s(a,x,c)$ is steeper than $y(a,x)$ with respect to a. The

Figure 4.1: The decision to apply elite institutions pre- and post-reform



Notes: The ability is denoted as a, and individual, family, and regional characteristics are denoted as x. All the total costs are denoted as c. The earning function, $y(a, x)$, represents if a student decides to stay in her own region. If students decide to apply to elite institutions, earning function is $y^s(a, x, c)$. The threshold level is denoted as $a*$, where if a student's ability is higher, she moves to the major city; if not, she stays in her region. The threshold changed based on a different case. Part (a) refers to the pre-reform case, part (b) refers if reform affected all students irrespective of their ability, part(c) refers if reform affected only higher ability ones, and part(d) refers to students those lives in major cities.

equilibrium is shown in Figure 4.1(a), where the student finds an optimal threshold: if their ability is higher than $a*$, they move to the major city; if their ability is lower than a^*, they stay in their region.

In the post-reform conditions, assuming that the USE reallocates talent and knowledge more efficiently, the costs c decrease and expected earnings increase. Figure 4.1(b) demonstrates that the net expected earnings increase, causing the curve y^s to shift upward, and the new threshold ability level a^* moves to the left, to the level of a^{**}. This indicates that the pre-reform case required a higher ability level to apply and move to a major city than the post-reform case. In scenario (c), we consider that the post-reform period improved the conditions only for students with high ability. This is illustrated in Figure 4.1(c), where the y^s slope increases only after surpassing the level of a^{**}. In the last scenario, introducing USE may disadvantage students who live in major cities because they face more competition caused by new incomers moving to these cities because of the reform. In this case, the y^s curve shifts upward, and the new threshold ability level, a^*, moves to the right a^{***}, as shown in Figure 4.1(d).

To summarize, my research aims to investigate whether the introduction of USE increased the redistribution of skills and talent, as well as expected earnings. I account for heterogeneity in ability, regions, and locations by employing various specifications. I use the Russian Longitudinal Monitoring Survey (RLMS) data, consisting of 25 annual rounds from 1994 to 2020 (excluding 1997 and 1999). The dataset encompasses 32 states (*oblast*) and 7 federal districts, with a total of 13,790 individuals. Ultimately, my study seeks to contribute to the ongoing discourse on the impact of USE on the Russian education system and offer valuable insights into the potential advantages and disadvantages of this policy.

My model assesses the effect of the USE reform on the average wage using the difference-in-difference model:

$$lnearnings_{it} = \phi(t) + \gamma d_{ijt} + \beta d_{ijt} \times I(t \geq s) + X'_{ijt}\delta + \epsilon_{ijt} \tag{4.1}$$

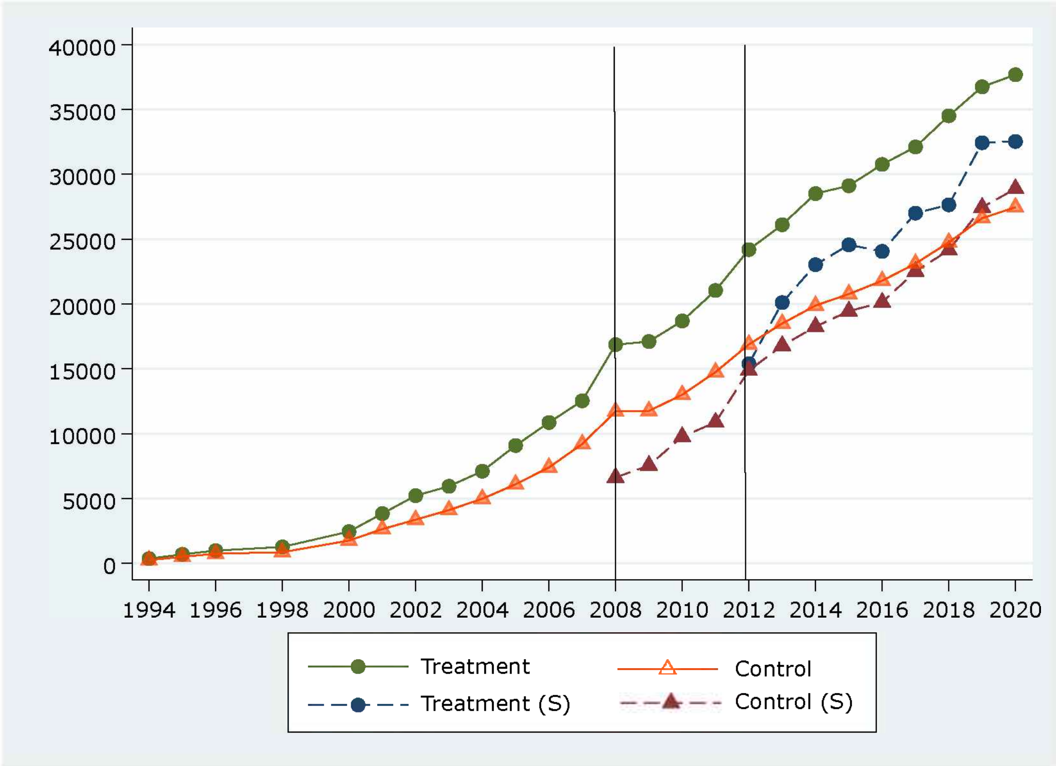
where $lnEarnings_{it}$ denotes the individual average monthly wage at time t. The dummy variable d_{ijt} equals one if the individual i in household j has a higher education diploma and equals zero otherwise. The function $I(.)$ indicates the occurrence of the implementation of the USE with s being the point in time in which the reform is fully implemented (s equals 2008 in the first wave and 2012 in the second wave). X_{ijt} is a vector of individual, household, and regional characteristics, and ϵ_{ijt} is a random error term. The treatment group consists of 3,576 individuals who possess higher education diplomas. This group was observed in each wave of interviews over the sample period, with a total of 67,837 person-wave observations during the reform-off period and 472 individuals observed during the reform-on period, with a total of 2,180 person-wave observations. The baseline control group, consisting of 8,146 individuals, was observed for a total of 219,527 person-wave observations during the reform-off period and 1,596 individuals for a total of 13,160 person-wave observations during the reform-on period.

I employ two different specifications for $\phi(t)$, which capture the time trends. In the first specification, I assume $\phi(t) = \tau_t$ and include a fully flexible set of time dummies that are the same to both the treatment and control groups. In the second specification, I assume $\phi(t) = \rho_0 + (\rho_1 + \rho_2 d_{ijt})t + [\rho_3 + \rho_4(t - s)]I(t \geq s)$, which is highly flexible as it allows for different intercepts (when $\gamma \neq 0$) as well as different linear trends for individuals in the treatment and control groups (when $\rho_1 \neq \rho_2$). Defining the treatment and control group, the USE is indeed a plausible source of exogenous variation that should affect the treatment group but not the control group. Also, I need to ensure that the groups are comparable in terms of observable and unobservable characteristics before the introduction of the treatment (making sure I take into account all the potential pre-treatment trends in the outcome variable). The 'treatment' group includes individuals who have a higher education; it consists of individuals who took the USE and graduated after the reform (thus were "treated" by the intervention). In comparison, the 'control' group consists of individuals completed their high-school education and applied to university before the introduction of the reform and thus were not affected by the reform. and have not been affected by the reform.

Figure 4.2 displays the average monthly earnings for both the treatment and baseline control groups over time (dashed lines restricted to the average salaries of individuals post-reform period). Prior to the reform, all groups exhibit similar patterns, thereby supporting the common trend assumption that we invoked when setting $\phi(t) = \tau_t$. The salaries of all groups share a similar trend both before and after the reform.

The trend in wages slowed down in 2008, likely due to the worldwide financial crisis. We are interested primarily in the group affected by the reform - individuals with higher education diplomas who finished high school after 2008 and attended

Figure 4.2: Average monthly salary in Russia, 1994-2020



Notes: 'Treatment' refers to the individual with a higher education diploma; it consists of individuals who took the USE and graduated after the reform (thus were "treated" by the intervention). 'Control' refers to the individual who has no higher education diploma. For both groups, the 'S' series (represented by dashed lines) denotes the average salary of individuals after the reform. The first vertical line indicates the introduction of the USE reform, at which point the Control (S) group enters the labor market. The second vertical line indicates the entry of the Treatment (S) group into the labor market after they graduate from higher education.

university [Treatment (S)]. Those who did not obtain a higher education degree enter the labor market with a lower salary. Our analysis shows that the post-reform treatment group has a slightly higher wage when they enter the labor market in 2012, despite their peers already having four years of experience. Furthermore, we observe that the wages of the treatment group increase very rapidly compared to both their peers and the older generation without a higher education diploma (Control). However, we also find that the average monthly salary of the treatment group has not yet caught up with the older generation with higher education diplomas, which can be explained by differences in experience. Overall, our findings suggest that the USE reform has had a positive impact on the wages of individuals with higher education diplomas, as evidenced by their higher salaries and more rapid wage growth compared to their peers and those without a higher education degree.

Table 4.1 displays the summary statistics of the primary explanatory variables before (reform-off) and after (reform-on) the USE reform across treatment and control groups. The p-values of a pre-reform balance test are presented in column 5. The results indicate that both treatment and control groups exhibit similar distributions in terms of individual and family characteristics, as well as regional and federal district features. The data show that more females than males have a higher education diploma, with the proportion slightly higher after the implementation of the reform. The parental education level is also in line with this finding, as the share of fathers' education decreased from 18% to 15%, while that of mothers' education increased from 21% to 27% in the treatment group. As expected, parental education plays a critical role in individual educational choices, and their proportions are lower among the control groups.

Regardless of whether the individuals were in the control or treatment group or subject to a reform off/on, the average household size was approximately 3.3 individuals. Although there were similarities in individual and household characteristics, significant differences were observed between groups in terms of location and federal districts. Notably, the proportion of individuals with higher education degrees decreased substantially in Moscow and Saint-Petersburg following the USE reform, whereas other major cities saw significant increases. Moreover, small cities and towns experienced a slight reduction in the proportion of higher education degrees, while rural areas saw slight increases. When examining the impact of the USE reform on federal districts, the North Caucasus, Southern, and Siberian federal districts saw significant increases in the proportion of individuals with university degrees in the overall population.

The baseline model identification might be plagued by selection bias. The outcome of USE reform might not be comparable between selected individuals. The use of Propensity Score Matching (PSM) addresses the bias that may result in the estimation of the effect of treatment on those outcomes. For this reason, I employ a two-step PSM approach that considers a wide range of observable characteristics and

Table 4.1: Descriptive Statistics for Expected Return on Higher Education

	Control group		Treatment group		Balancing test
	Reform-off	Reform-on	Reform-off	Reform-on	p-value
Individual characteristics					
Age	47.71	21.06	45.42	24.51	0.903
Female	0.570	0.511	0.623	0.662	0.003
Married	0.647	0.270	0.712	0.452	0.269
Russian ethnicity	0.842	0.850	0.876	0.884	0.001
Born elsewhere	0.538	0.290	0.510	0.302	0.003
Household demographics					
Mather education	0.083	0.212	0.212	0.270	0.000
Father education	0.054	0.105	0.183	0.150	0.000
Household size	3.304	3.879	3.225	3.394	0.008
Number of children, 0-6	0.226	0.271	0.265	0.266	0.000
Number of children, 6-17	0.337	0.320	0.352	0.139	0.821
Locations					
Moscow	0.064	0.073	0.136	0.105	0.002
St. Petersburg	0.028	0.024	0.046	0.024	0.001
Other major cities	0.273	0.296	0.398	0.461	0.003
Small cities and towns	0.341	0.304	0.296	0.270	0.003
Rural areas	0.293	0.259	0.124	0.140	0.002
Federal districts					
North and North western	0.094	0.086	0.103	0.069	0.002
Central	0.271	0.286	0.359	0.352	0.000
Volga	0.238	0.182	0.196	0.194	0.035
North Caucasus	0.038	0.068	0.021	0.035	0.468
Southern	0.104	0.122	0.097	0.135	0.000
Ural	0.086	0.080	0.077	0.057	0.000
Siberia	0.139	0.150	0.133	0.143	0.049
Number of observation	219,527	13,160	67,837	2,180	
Number of individual	8146	1596	3576	472	
Monthly Salary	13296.04	21764.24	21278.26	28418.57	
Number of observation	102,951	4,911	43,828	1,484	

Notes: The numbers are mean by group (Treatment and Control) and period (reform-off = 1994-2007, reform-on = 2008-2020). 'Treatment' refers to the individual who has higher education diploma and 'Control' refers to the individual who has no higher education diploma.

adjusts for the differences identified in Table 4.1. This methodology combines the difference-in-difference estimator with a matching technique that pairs each treated individual with a subset of individuals in the untreated group with the most similar observable characteristics. I estimated the average treatment effect (ATE), average treatment on the treated effect (ATT), and the average treatment on the untreated wage earnings effect (ATU) of having a higher education diploma both pre- and post-reform periods.

The necessary weights for this process are calculated using PSM in the first stage, while the second stage is estimated using weighted least squares and a complete set of time dummies as recommended by Blundell & Dias (2009). Propensity scores are derived from a logistic regression model, where the dependent variable is the treatment indicator variable multiplied by an indicator variable for the post-reform period, and the independent variables are all the covariates listed in Table 4.1. In order to create a balanced and comparable sample between the treatment and control groups, it is necessary to employ a matching strategy that pairs each treated individual, defined as one who holds a higher education degree and has been impacted by the reform, with a subset of non-treated individuals who lack a higher education degree and have not been subjected to the reform, while also ensuring that these non-treated individuals possess similar observable characteristics. We employ a kernel-matching technique with an Epanechnikov kernel and a bandwidth of 0.05. We conducted sensitivity analyses by varying the kernel and bandwidth, and our results remain robust to these choices.

In my analysis, selecting an appropriate control group is crucial due to the nationwide roll-out of the reform occurring simultaneously. The selection of control groups could be problematic as they may be chosen based on the endogenous outcome. To mitigate this issue, I considered an alternative plausible comparison group. This comparison group is created by limiting the sample to individuals aged between 18 and 30 years. Since individuals affected by the reform were 30 years old in 2020, the pre-reform groups were also restricted to this age range. Table 4.5 presents the results obtained using this comparison group, confirming my baseline estimates and indicating that the strongest impact of USE reform has been seen in major cities, small cities, and towns. Looking ahead to our findings, we do not detect any signs of potential bias resulting from differences in observed characteristics between the groups.

4.4 Results

In this section, I present the estimated impact of the USE reform on the average monthly salaries. Table 4.2 shows the results for the full sample and a restricted sample with specific regions. Panel A displays the results from a linear specification model in which we imposed a fully flexible set of time dummy variables that are common to both treatment and control groups. Panel B presents the estimation results from the specification that includes group-specific time trends. All standard errors are robust to arbitrary forms of heteroscedasticity, and all are clustered at the population center level. Panel C represents the propensity score matching estimates obtained with the two-step Blundell-Costa Dias procedure (Blundell & Dias 2009).

As mentioned above, I divided our participants into two groups: the 'Treatment' group and the 'Control' group. The Treatment group includes students who took the USE in 2008 or later, received their university degree, and were fully affected by the reform. These are the students who took the exam under the new system and experienced the changes brought about by the reform. On the other hand, the Control group consists of students with no higher education diploma and were not affected by the reform. I also focused on comparisons between the high-school graduates without university degree that graduated prior the reform and after the reform was implemented. I focused on estimating the common trends between the groups, and found that the difference in average monthly salary between the high-school educated students was 5.5%. The reform contributed 8% more on average to the treatment group.

All three panels of Table 4.2 have similar results; the differences in magnitude of the effects are very minor. The study confirms that having a higher education diploma leads to higher earnings compared to the high-school only educated peers, and that the younger generation has an advantage in the labor market over the older generation. This advantage may be due in part to the reform, which helped younger people acquire new skills. The results of PSM in panel C of Table 4.2 show again that the reform affected both levels of education positively.

The reason why the younger generation earns more on average than the older generation can be attributed to a couple of factors. One of the main reasons is their technological skills. Growing up in a world where technology plays a significant role has made the younger generation more adept at using it. This, in turn, has made them more valuable in tech-related fields and has allowed them to earn higher salaries and obtain more advanced positions. Another factor contributing to the higher earnings of the younger generation is the career stage at which they typically start working. They tend to begin their careers during periods of economic growth, when job opportunities and potential for advancement are at their highest. This head start in career growth and earning potential gives them an advantage over older generations who may have entered the workforce during times of economic decline.

Moreover, the economic conditions that prevailed during the time when the younger generation entered the workforce were markedly different from those of previous generations. The young generation entered the workforce during a period of sustained economic growth, which provided them with more job opportunities and higher salaries. Finally, the diversity of the young generation has played a role in their success. With a larger share of women and minorities who have historically been underpaid, there is a greater push for equal pay and opportunities for these groups. This has contributed to a more level playing field for young people entering the workforce today.

One of the main intentions of introducing USE was to reduce university admission exam costs, especially in prestigious universities in Moscow, St. Petersburg, and other major cities. All regions, including St. Petersburg, are in line with the full sample. *Treatment pre-reform* group earns highest, which expected as they are more experienced and spent decades in the labor market, it follows by *Treatment post-reform* group, *Control post-reform* group and *Control pre-reform* group. The exception is Moscow, where results show that reform had more significant impact on the average salary of the young generation (Table-4.2, panel-A). *Treatment post-reform* group even earns more than *Treatment pre-reform* in small cities and towns. All outcomes are similar when we account group specific linear trends. The corre-

Table 4.2: Expected Return on Higher Education

	Full sample	Moscow	St.Petersburg	Other Major cities	Small cities and towns	Rural areas
Panel A:						
Post-reform	0.055***	0.004	0.086	0.036**	0.059***	0.047**
	(0.008)	(0.008)	(0.072)	(0.011)	(0.016)	(0.018)
Treatment	0.108***	0.110***	0.073***	0.094***	0.096***	0.093***
	(0.007)	(0.001)	(0.021)	(0.012)	(0.007)	(0.011)
Post$\times$ *Treatment*	-0.083***	0.024*	-0.051***	-0.076**	-0.057*	-0.061
	(0.022)	(0.008)	(0.013)	(0.024)	(0.028)	(0.041)
N	102212	7373	3063	33208	36792	21805
R^2	0.807	0.818	0.799	0.795	0.818	0.800
Panel B:						
Post-reform	0.116***	0.113**	0.187*	0.126***	0.111***	0.103***
	(0.008)	(0.008)	(0.074)	(0.013)	(0.014)	(0.016)
Treatment	0.078***	0.079***	0.053*	0.067***	0.075***	0.067***
	(0.005)	(0.000)	(0.021)	(0.008)	(0.006)	(0.007)
Post$\times$ *Treatment*	-0.062**	-0.002	-0.053***	-0.068*	-0.046	-0.030
	(0.021)	(0.008)	(0.140)	(0.002)	(0.029)	(0.040)
N	102212	7373	3063	33208	36792	21805
R^2	0.796	0.799	0.779	0.776	0.809	0.786
Panel C:						
Post-reform	0.234***	0.197***	0.169	0.118***	0.202***	0.238***
	(0.027)	(0.000)	(0.231)	(0.030)	(0.032)	(0.054)
Treatment	0.133***	0.291***	0.218***	0.071**	0.100*	0.090**
	(0.045)	(0.001)	(0.101)	(0.029)	(0.046)	(0.045)
Post$\times$ *Treatment*	-0.101***	0.094***	-0.287***	-0.046***	-0.103	-0.148**
	(0.034)	(0.001)	(0.114)	(0.008)	(0.065)	(0.061)
N	101949	6992	2326	32261	35823	20426
R^2	0.667	0.661	0.614	0.771	0.628	0.696

Notes: The dependent variable is the logarithm monthly salary of individuals. Panel-A, columns represents the results from a linear specification model in which imposed a fully flexible set of time dummy variables common to treatment and control groups, while Panel-B represents estimation from specification that includes group-specific time trends. Panel-C represents PSM-based estimates that are obtained from a two-step procedure. *Post-reform* refers to the results of 'Control' "S" group - people who have no higher education diploma and were affected by the reform, *Treatment* refers to results of 'Treatment' group - people who have higher education diploma and were not affected by the reform, *sum of all* refers results of 'Treatment' "S" group - people who have higher education diploma and were affected by the reform. All results compare average monthly salary growth with the 'Control' group - people who have no higher education diploma and were unaffected by the reform. All standard errors are robust to arbitrary forms of heteroscedasticity, and all are clustered at the population centre level. Standard errors are in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

sponding effect estimates are reported in the next panel of (Table-4.2, panel-B). The *post-reform* gain statistical significance for Moscow, reform did play positive impact. The impact is very significant for the individual from other cities (major and small) and towns.

The proposed framework posits that the returns on education can be influenced by a range of factors, including individual and family characteristics, as well as the prestige of the university from which the individual graduates. Crucially, the choice of university is largely predicated on an individual's academic achievements in secondary school, high school, and regional and individual/family characteristics. However, it is important to recognize that families from small towns and rural areas often have less financial resources compared to their counterparts in major cities. This may limit their ability to provide adequate support for their children to move to urban areas for higher education, unless they are admitted to tuition-free education programs. The efficacy of the USE reform in promoting social mobility remains a topic for future research as it is still early to observe its full impact. Furthermore, the trend of studying abroad, particularly in Europe and the US, has become increasingly prevalent, which may lead to a decline in the quality of students who choose to attend domestic institutions.

The further results from different exercises to check the robustness of estimates are reported in Table-4.3. We ask whether reform had a heterogeneous effect on the educational return along a number of observable characteristics. In particular, we investigate the possibility of differential responses by gender, ethnicity, household salary, and parental education. The main interest results are in line with the baseline model. There is significant heterogeneity by gender, marital status, parental education, and the number of children in the household, irrespective of other specifications.

The primary findings align with the baseline outcomes, demonstrating a substantial effect of the reform among individuals residing in small cities and towns. Moreover, the impact of the reform is particularly noteworthy among residents of Moscow. Furthermore, the reform contributed significantly more to those individuals who earned their university diplomas after the implementation of the reform.

Table 4.3: Heterogeneous Effects of Expected Return on Higher Education

	Full sample	Moscow	St.Petersburg	Other Major cities	Small cities and towns	Rural areas
Reform outcome:						
Post-reform	0.078***	-0.005	0.057	0.065***	0.128***	0.046
	(0.014)	(0.006)	(0.081)	(0.017)	(0.023)	(0.028)
Treatment	0.191***	0.164***	0.051**	0.173***	0.187***	0.193***
	(0.011)	(0.001)	(0.023)	(0.017)	(0.012)	(0.020)
Post×Treatment	-0.121***	0.062***	-0.046***	-0.077**	-0.155***	-0.114***
	(0.023)	(0.004)	(0.014)	(0.035)	(0.040)	(0.042)
Individual characteristics:						
Age	0.030***	0.028***	0.024***	0.036***	0.025***	0.016***
	(0.002)	(0.000)	(0.006)	(0.002)	(0.003)	(0.005)
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.307***	-0.198***	-0.158***	-0.306***	-0.289***	-0.239***
	(0.011)	(0.001)	(0.023)	(0.015)	(0.013)	(0.017)
Married	0.017**	0.024**	-0.016	-0.010	-0.005	0.048***
	(0.008)	(0.003)	(0.024)	(0.016)	(0.015)	(0.016)
Russian ethnicity	0.038*	-0.010	0.056	-0.025	0.010	-0.038
	(0.021)	(0.006)	(0.039)	(0.018)	(0.051)	(0.041)
Born elsewhere	0.023**	0.016**	0.011	0.010	0.062*	0.050***
	(0.009)	(0.002)	(0.022)	(0.012)	(0.033)	(0.016)
Household demographics:						
Mather education	0.064***	0.023***	0.055*	0.124**	0.005	0.006
	(0.011)	(0.000)	(0.030)	(0.101)	(0.013)	(0.022)
Father education	0.067***	0.050***	0.004	0.039	0.053**	0.094***
	(0.013)	(0.000)	(0.034)	(0.023)	(0.023)	(0.024)
Household size	0.007	0.014***	0.015*	0.005	-0.007	0.002
	(0.005)	(0.000)	(0.009)	(0.006)	(0.009)	(0.005)
Child, 0-6	-0.052***	-0.089***	-0.114***	-0.087***	-0.015	-0.022*
	(0.009)	(0.004)	(0.027)	(0.013)	(0.016)	(0.011)
Child, 6-17	-0.000	-0.018**	-0.014	-0.000	0.024**	-0.017*
	(0.006)	(0.002)	(0.021)	(0.009)	(0.011)	(0.010)
Constant	5.511***	6.307***	7.459***	5.427***	5.584***	5.351***
	(0.128)	(0.005)	(0.139)	(0.080)	(0.129)	(0.120)
N	102212	7373	3063	33208	36792	21805
R^2	0.763	0.784	0.780	0.770	0.761	0.750

Notes: The dependent variable is logarithm of monthly salary of individual. The first column represents full sample with controlling locations, federal districts, years and trends dummies. The following columns represent result for sample that restricted to specific locations. *Post-reform* refers to the results of 'Control' "S" group - people who have no higher education diploma and were affected by the reform, *Treatment* refers to results of 'Treatment' group - people who have higher education diploma and were not affected by the reform, *sum of all* refers results of 'Treatment' "S" group - people who have higher education diploma and were affected by the reform. All results compare average monthly salary growth with the 'Control' group - people who have no higher education diploma and were unaffected by the reform. All standard errors are robust to arbitrary forms of heteroscedasticity, and all are clustered at the population centre level. The control variables are listed in Table 1. Standard errors are in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The results also suggest that the reform did not benefit individuals who were affected by the reform but did not obtain a diploma in Moscow. However, this group earns higher than the older generation without a diploma in major cities, small cities and towns.

Females earn 30 % less than males; results are similar irrespective of different specifications and locations. That might be explained by occupational segregation, where females are more likely to work in low-paying occupations and industries that are traditionally female-dominated, such as education, healthcare, and social services. Another reason is that females take time off to care for children or work part-time to balance work and family responsibilities. Last but not least, women are underrepresented in senior management positions and on corporate boards in Russia, which can limit their opportunities for advancement and high-paying roles.

Married individuals earn more on average, where being married may provide greater motivation for individuals to succeed in their careers, as they may be more focused on providing for their family and securing their financial future. Russian ethnicity significantly matters in full sample specification, but it is not significant once we restricted sample to the participial location. Born elsewhere has signification contribution the salary, that might indicates that higher motivation of individual, hard working, might have higher experiences.

The influence of parental education on an individual's salary is a well-documented phenomenon in the academic literature. Our findings confirm that having a parent with a higher education diploma significantly increases an individual's average salary by 6% for each parent. Moreover, if both parents have a higher education diploma, the individual is expected to earn 13% more than their peers who do not have such a family background. However, the impact of parental education is not uniform across different regions in Russia. Interestingly, the impact of parents' education varies across regions in Russia. In Moscow, both parents' education levels matter, with the father's education playing an even more significant role. On the other hand, the mother's education level is more critical in St. Petersburg and other major cities. In contrast, the father's education level plays a more significant role in determining an individual's salary in small cities and towns.

On the other hand, having a child under six years old in the household reduces the average salary due to parental leave or a part-time job for baby care. This impact is more severe in major cities, where the cost of living is higher, and the competition in the job market is more intense. The gender of the parent taking care of the child also affects the salary. Women are more likely to take time off work or work part-time to care for their children, resulting in a more significant negative impact on their earnings. Overall, our findings suggest that parental education and childcare policies are essential factors that shape an individual's earning potential and should be considered in designing policies to promote social mobility and gender equality.

As part of our robustness analysis, we sought to ensure the consistency and reliability of our findings. To that end, we conducted additional tests using average monthly income as the dependent variable, which is presented in the Appendix (Table-4.4). The results are provided for the full sample and are controlled for location and federal districts, and clustered at the population center level. The subsequent columns provide outcomes restricted to specific locations. The linear model is presented without control variables in Panel-A and with control variables in Panel-B. The model with group-specific time trends is presented without control variables in Panel-C and with control variables in Panel-D. Our analysis suggests that all four different specifications and six different restricted samplings have yielded similar results. We observed that the trend in income differs from that of salaries, with the highest income belonging to the *Treatment pre-reform* group, followed by the *Control pre-reform* group, the *Treatment post-reform* group, and finally the *Control post-reform* group.

The average income is higher than salary for different reasons. The older generation has additional sources of income, such as rental income, investment income, or business income, which contribute to their total income. Bonuses and commissions are correlated with experience, where age matters. Lastly, older generations might have more self-employment income, and they have the potential to earn more than salaried employees, depending on their industry and level of success. These findings provide additional support for the robustness of our results, which indicate that our earlier findings are reliable and can be generalized to a wider population.

4.5 Conclusion

This study provides evidence on the impact of the unified state exam (USE) on education returns in Russia. The USE is a national level exam for university admissions that aims to increase accessibility and efficiency by reducing admission costs and allowing students from smaller cities, towns, and rural areas to apply to universities in other locations. The previous admission system had multiple institution-specific exams, higher costs, and limited university choices. The USE eliminates these barriers and maximizes productivity, leading to higher education returns.

According to our baseline model, the reform has had a positive impact on education returns, and the effect is statistically significant. The impact is more prominent in Moscow and St. Petersburg, where most elite schools are located. One reason for this could be the increased mobility of talented individuals from small cities, towns, and rural areas to bigger cities, where salaries are higher. Late entry into the labor market does not put those who spent four years at university at a disadvantage, as they start earning slightly more than their peers without higher education degrees, despite entering the labor market four years later. Their salaries grow rapidly, and after spending one or two years at work, they even earn more than their older counterparts without higher education qualifications.

These findings are robust and consistent across various methods of analysis, different control groups, and model specifications. Notably, factors such as family background, regional and federal district location have a significant impact on an individual's salary in Russia. Additionally, an individual's gender, marital status, ethnicity, and parental education also significantly influence their earnings. Women tend to earn less than men, married individuals earn more than unmarried, and having a parent with a higher education degree increases an individual's average salary. It is worth noting that the influence of parental education varies across different regions of Russia. In Moscow and smaller cities and towns, the father's education plays a more significant role, while in St. Petersburg and other major cities, the mother's education level has a more significant impact.

While this study sheds light on the positive impact of the USE on education return, there are still many aspects that need to be explored in future research. For example, it is the investigation whether the USE had any unintended consequences on social development and family life lies ahead. Furthermore, it is important to explore whether the introduction of the USE has encouraged secondary schools to improve their teaching quality or discouraged students from participating effectively in class due to a narrow focus on exam preparation. Another question to consider is whether the USE has caused students to rely only on private tutors for exam preparation, leading to unequal access to education. Last, it is worth examining whether the USE fully measures all the skills and abilities of students, including social, interpersonal, and presentational skills, which are vital for success in many professions. Future research can delve into these questions and provide a more comprehensive understanding of the impact of the USE on education and society as a whole.

References

Ampilogov, A., I. Prakhov, & M. Yudkevich (2013): "One or Many? Using the New Opportunities of the Unified State Exam in Russian University Admissions." *HSE Working paper 12/EDU/2013*, National Research University Higher School of Economics, Moscow.

Avery, C., S. Lee, & A. E. Roth (2014): "College Admissions as Non-Price Competition: The Case of South Korea." *NBER Working Paper 20774*, National Bureau of Economic Research.

Blundell, R. & M. C. Dias (2009): "Alternative approaches to evaluation in empirical microeconomics." *Journal of Human Resources* **44(3)**: pp. 565–640.

Borisov, G. & C. A. Pissarides (2016): "The intergenerational transmission of human capital and earnings in contemporary Russia." *Iza discussion paper*, Institute of Labor Economics.

Chade, H., G. Lewis, & L. Smith (2014): "Student Portfolios and the College Admissions Problem." *The Review of Economic Studies* **81(3-288)**: pp. 971–1002.

Che, Y.-K. & Y. Koh (2016): "Decentralized College Admissions." *Journal of Political Economy* **124(5)**: pp. 1295–1338.

Fedotova, O. & O. Chigisheva (2010): "Restructuring the Governance and Management Structures of Higher Education in Russia." In D. Johnson (editor), "Politics, Modernisation and Educational Reform in Russia: From Past to Present," Oxford Studies in Comparative Education, chapter 5, pp. 81–98. Symposium Books: Oxford.

Francesconi, M., F. Slonimczyk, & A. Yurko (2019): "Democratizing Access to Higher Education in Russia: The Consequences of the Unified State Exam Reform." *European Economic Review* **117(C)**: pp. 56–82.

Goodman, J., O. Gurantz, & J. Smith (2020): "Take Two! SAT Retaking and College Enrollment Gaps." *American Economic Journal: Economic Policy* **12(2)**: pp. 115–158.

Gounko, T. & W. Smale (2007): "Modernization of Russian Higher Education: Exploring Paths of Influence." *Compare: A Journal of Comparative and International Education* **37(4)**: pp. 533–548.

Hafalir, I. E., R. Hakimov, D. Kubler, & M. Kurino (2018): "College Admissions with Entrance Exams: Centralized versus Decentralized." *Journal of Economic Theory* **176(C)**: pp. 886–934.

Kerr, S. P., T. Pekkarinen, M. Sarvimaki, & R. Uusitalo (2020): "Postsecondary Education and Information on Labor Market Prospects: A Randomized Field Experiment." *Labour Economics* **66(C)**: pp. 1–14 (art. 101888).

Khavenson, T. & A. Solovyova (2014): "Studying the Relation between the Unified State Exam Points and Higher Education Performance." *Voprosy obrazovaniya*

/ *Educational Studies Moscow* **2014(1)**: pp. 176–199.

Luk'yanova, E. (2012): "Russian Educational Reform and the Introduction of the Unified State Exam. A View from the Provinces." *Europe-Asia Studies* **64(10)**: pp. 1893–1910.

Minina, E. (2010): "The Unified National Test for Student Admission to Higher Education in Russia: A Pillar of Modernisation?" In D. Johnson (editor), "Politics, Modernisation and Educational Reform in Russia: From Past to Present," Oxford Studies in Comparative Education, chapter 7, pp. 121–136. Symposium Books: Oxford.

Prakhov, I. (2021): "The Determinants of Expected Returns on Higher Education in Russia: A Human Capital Theory Approach." *Higher Education Quarterly* **75(2)**: pp. 302–320.

Prakhov, I. & M. Yudkevich (2019): "University Admission in Russia: Do the Wealthier Benefit from Standardized Exams?" *International Journal of Educational Development* **65(C)**: pp. 98–105.

RMES (2014): "Russian Ministry of Education and Science, Unified State Examination: Final State Certification (in Russian)." *Regulation order, july 30, 2014, n875*, Moscow: Ministry of Education and Science.

RMES (2018): "Russian Ministry of Education and Science, Unified State Examination: Final State Certification (in Russian)." *Regulation order, september 18, 2018, n1021*, Moscow: Ministry of Education and Science.

RMES (2019): "Russian Ministry of Education and Science, Unified State Examination: Final State Certification (in Russian)." *Regulation order, september 17, 2019, n1094*, Moscow: Ministry of Education and Science.

Rosstat (2021): "In: Statistical Yearbook of Russia." *Statistical yearbook*, Rosstat, Moscow.

4.A Supplementary Statistics

Table 4.4: Total Income as Expected Return on Higher Education

	Full sample	Moscow	St.Petersburg	Other Major cities	Small cities and towns	Rural areas
Panel A:						
Post-reform	-1.654***	-1.755***	-1.610***	-1.560***	-1.791***	-1.696***
	(0.049)	(0.016)	(0.137)	(0.088)	(0.097)	(0.088)
Treatment	0.347***	0.272***	0.183***	0.254***	0.320***	0.402***
	(0.018)	(0.003)	(0.062)	(0.026)	(0.027)	(0.035)
Post×Treatment	1.257***	1.776***	0.309	1.270***	1.269***	1.143***
	(0.098)	(0.021)	(0.402)	(0.134)	(0.200)	(0.205)
N	234991	17432	6309	69467	78249	63600
R^2	0.477	0.433	0.467	0.470	0.488	0.476
Panel B:						
Post-reform	-0.945***	-1.123***	-1.088***	-0.903***	-0.978***	-0.918***
	(0.048)	(0.017)	(0.151)	(0.078)	(0.089)	(0.101)
Treatment	0.360***	0.308***	0.157**	0.316***	0.392***	0.496***
	(0.018)	(0.004)	(0.064)	(0.020)	(0.031)	(0.040)
Post×Treatment	1.115***	1.653***	0.349	1.110***	1.068***	0.972***
	(0.102)	(0.019)	(0.400)	(0.122)	(0.218)	(0.196)
N	234991	17432	6309	69467	78249	63600
R^2	0.489	0.440	0.479	0.481	0.502	0.486
Panel C:						
Post-reform	-1.829***	-1.807***	-2.170***	-1.786***	-2.021***	-1.757***
	(0.074)	(0.032)	(0.265)	(0.144)	(0.142)	(0.127)
Treatment	0.912***	0.592***	0.294*	0.723***	0.842***	0.964***
	(0.040)	(0.005)	(0.164)	(0.052)	(0.062)	(0.109)
Post×Treatment	1.290***	1.797***	0.075	1.273***	1.262***	1.154***
	(0.100)	(0.035)	(0.416)	(0.128)	(0.191)	(0.226)
N	234991	17432	6309	69467	78249	63600
R^2	0.474	0.431	0.465	0.467	0.485	0.469
Panel D:						
Post-reform	-1.146***	-1.191***	-1.635***	-1.128***	-1.220***	-1.045***
	(0.068)	(0.033)	(0.275)	(0.128)	(0.129)	(0.135)
Treatment	0.793***	0.461***	0.077	0.642***	0.776***	0.972***
	(0.043)	(0.006)	(0.166)	(0.055)	(0.062)	(0.113)
Post×Treatment	1.112***	1.617***	0.014	1.070***	1.036***	0.969***
	(0.105)	(0.032)	(0.413)	(0.121)	(0.205)	(0.214)
N	234991	17432	6309	69467	78249	63600
R^2	0.484	0.438	0.476	0.476	0.496	0.477

Notes: The dependent variable is the logarithm of the average monthly income. The first column represents the full sample, controlling the different locations and federal districts. The following columns represent results restricted to specific locations. The results from the linear specification model without control variables in Panel-A, with control variables in Panel-B. The results from the model specification include group-specific time trends without control variables in Panel-C, with control variables in Panel-D. *Post-reform* refers to the results of 'Control' "S" group - people who have no higher education diploma and were affected by the reform, *Treatment* refers to results of 'Treatment' group - people who have higher education diploma and were not affected by the reform, *sum of all* refers results of 'Treatment' "S" group - people who have higher education diploma and were affected by the reform. All results compare average monthly salary growth with the 'Control' group - people who have no higher education diploma and were unaffected by the reform. All standard errors are robust to arbitrary forms of heteroscedasticity, and all are clustered at the population centre level. The control variables are listed in Table 1. Standard errors are in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.5: Sensitivity check - Expected Return on Higher Education

	Full sample	Moscow	St.Petersburg	Other Major cities	Small cities and towns	Rural areas
Post-reform	0.040**	-0.002	0.215	0.024 **	0.061*	0.030
	(0.015)	(0.004)	(0.119)	(0.009)	(0.027)	(0.035)
Treatment	0.112***	0.133***	0.108	0.097**	0.116***	0.065*
	(0.015)	(0.001)	(0.044)	(0.024)	(0.021)	(0.029)
$Post \times Treatment$	-0.050*	0.055*	-0.062	-0.022***	-0.053*	-0.0320**
	(0.024)	(0.026)	(0.072)	(0.003)	(0.025)	(0.011)
N	22134	1522	640	7866	7644	4465
R^2	0.752	0.712	0.763	0.742	0.772	0.758

Notes: The dependent variable is the logarithm of the average monthly salary. The first column represents the full sample, controlling the different locations and federal districts. The following columns represent results restricted to specific locations. The pre-reform groups were restricted to 18-30. *Post-reform* refers to the results of ‘Control’ “S” group - people who have no higher education diploma and were affected by the reform, *Treatment* refers to results of ‘Treatment’ group - people who have higher education diploma and were not affected by the reform, *sum of all* refers results of ‘Treatment’ “S” group - people who have higher education diploma and were affected by the reform. All results compare average monthly salary growth with ‘Control’ group - people who have no higher education diploma and were unaffected by the reform. All standard errors are robust to arbitrary forms of heteroscedasticity, and all are clustered at the population centre level. The control variables are listed in Table 1. Standard errors are in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Response to Comments from Reviewers and Committee

Three Essays on the Economics of Education

Mgr. Olesia Zeynalova

Small defense on February 15, 2023

Revision completed: March 31, 2023

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*; my response is in roman type. To keep the response shorter, I have deliberately chosen only passages of reports that call for response or warrant revision.

Response to Comments from dr. Jan Janku

First paper. *(..) Despite the fact that I do not have many fundamental comments on the published meta-analysis and I stress again that I consider it very carefully prepared, I raise below a few (rather debatable) questions:*

(i.) *Since the primary studies included in the meta-analysis used different transformations of the variables and different specifications (linear, log, semi-log specifications of the demand function), the author(s) convert the collected estimates into partial correlation coefficients (PCC). This is a standard procedure, but can still raise some controversy (Reed 2020). As Reed (2020) suggests,* *meta-analysts should be cautious in extrapolating results based on PCC analyses. Analyses using PCCs will produce similar results only when the inverses of the standard errors of the PCCs are closely related (high correlation) to those of the estimated effects. Thus, it is up for debate whether meta-analyses of this type (using PCC) should be accompanied by a simple correlation analysis as suggested by Reed (2020). I should however also add that in the paper under review, this concern of mine is alleviated by the robustness check that is in the third panel of Table 2.7. If I understand correctly, the author(s) here use only a subset of the specification (log-log) and this specification is not transformed into PCC. As a consequence, the resulting coefficients can be interpreted as elasticities. Since the direction and significance of the estimates are not very different from regressions using PCC, this shows that PCCs are used correctly.*

This is a very good point. In fact, standard errors of PCCs are not only "closely related" to those of the estimated effects, they are correlated by definition. Yes, the partial correlation in meta-analyses is used as a work-a-round and one should always prefer to use the effect sizes. When we have submitted the manuscript to OBES, a bright referee told us to do a robustness check with a subset of estimates that were directly comparable – that is, those reported in terms of elasticities. If the baseline PCC analysis was in line with the analysis of the elasticities subset, we should feel comfortable using PCCs. This referee's advice remains the rule of thumb regarding PCCs, and probably the best we can do in practice. Without PCCs the potential for meta-analysis in economics would shrink substantially. Also note that if one uses (any kind of) correlations, these correlations are automatically perceived as non-causal. The outcome of the paper is, nevertheless, a zero effect. In such case, ordinality of PCCs is sufficient and serves well the purpose.

(ii.) *My other rather minor question is why the author(s) did not use any of the nonlinear techniques described in the recent literature. These methods include, for example: Top 10 Method by Stanley* et al. *(2010), Weighted Average of Adequately Powered (WAAP) by Ioannidis* et al. *(2017), Selection Model by Andrews & Kasy (2019), Stem-based Method by Furukawa (2020) and Kinked Method by Bom & Rachinger (2019). These methods in contrast to the linear ones do not focus on quantifying publication bias itself but rather on estimating the effect beyond bias (the true effect). Presumably, the author(s) could use this to further test the true effect, which turns out to be statistically insignificant in most of their regressions - thus contradicting most published primary studies.*

The paper was accepted for publication in early 2018 so most of the toolbox you mention was not available at the time of the submission to the journal. Also, both Top10 method and WAAP perform well only if there is no heterogeneity in the data (which rarely happens especially in economics research). Both are suggestive but also overly simplistic. You are correct though, the funnel-based tests do not work properly when the correlation mentioned in the previous point is a given. But neither IV specification nor proxy specification in Table 2.3 assume the linear (nor any other, for that matter) relationship between PCCs and their standard errors. They either instrument or proxy the number of observations for the standard error. I ran the numbers using non-linear techniques as well and the results are summarized in the following Table 4.6. Note that WAAP by Ioannidis *et al.* (2017) could not be estimated due to insufficient observations with adequate power. Bottom-line: all the novel methods for publication bias correction support our original conclusions.

Table 4.6: Non-linear methods suggested by the referee

	Top10	Stem-based	Kinked-meta	Selection model
Publication bias			-2.187***	P = 0.844
			(0.169)	(0.459)
Effect beyond bias	-0.005	-0.032	0.021***	0.007***
	(0.019)	(0.042)	(0.006)	(0.002)
Observations	442	442	442	442

Notes: Kinked-meta is endogenous kink model by Bom & Rachinger (2019), Stem model is by Furukawa (2020), selection is model by Andrews & Kasy (2019) using clustered SEs, P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). Standard errors, clustered at the study level, are reported in parentheses.

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

(iii.) *I also wonder what the results would be if the author(s) winsorize the elasticities obtained (at least at the 1 and 99 percentile). In fact, some of the very outlier values may bias the whole analysis. I do not have a very clear opinion on whether winsorization should be used in meta-analysis, but it would be at least interesting to compare notwinsorized and winsorized regressions. However, it is quite possible that the author(s) have winsorized elasticities and just do not explicitly state this in the text.*

We did not identify substantial outlying observations back in 2017. For demonstration, I replicate the results of Table 2.3 using winsorization at 1% level. From Table 4.7 we can see the effect sizes changed only at the third decimal place suggesting the winsorization is unnecessary. The results once again support the original conclusions of Chapter 2.

(iv.) *Finally, the author(s) should, in my opinion, comment somewhat more on the fact that their results differ from the previous quantitative synthesis by Gallet (2007). They should explain why the results are (arguably) different. This concerns, for example, the functional form of the demand function. Indeed, Gallet (2007) states that the elasticities differ according to the functional form used, whereas according to the author(s) these elasticities do not differ.*

Great question. The differences would be easy to identify if our samples perfectly overlapped but this is not the case. We use parts of the dataset by

Table 4.7: Funnel asymmetry tests using winsorized data

Panel A: Unweighted sample	OLS	IV	Proxy	Median
SE (publication bias)	-1.174***	-1.949***	-1.531***	-1.363***
	(0.37)	(0.42)	(0.27)	(0.50)
Constant (effect absent bias)	-0.058	0.017	-0.005	-0.035
	(0.06)	(0.04)	(0.04)	(0.07)
Observations	442	442	442	442
Panel B: Weighted sample	Precision		Study	
	WLS	IV	OLS	IV
SE (publication bias)	-1.783***	-2.325***	-1.072***	-1.820***
	(0.35)	(0.48)	(0.31)	(0.52)
Constant (effect absent bias)	0.001	0.027	-0.069**	0.017
	(0.02)	(0.02)	(0.03)	(0.04)

Notes: The table reports the results of the regression $PCC_{ij} = PCC_0 + \beta \cdot SE(PCC_{ij}) + \mu_{ij}$, where PCC_{ij} denotes i-th tuition elasticity of demand for higher education estimated in the j-th study and $SE(PCC_{ij})$ denotes its standard error. Panel A reports results for the whole sample of estimates, and Panel B reports the results for the whole sample of estimates weighted by precision or study. OLS = ordinary least squares. IV = the inverse of the square root of the number of observations is used as an instrument for the standard error. Proxy = the inverse of the square root of the number of observations is used as a proxy for the standard error. Median = only median estimates of the tuition elasticities reported in the studies are included. Study = model is weighted by the inverse of the number of estimates per study. Precision = model is weighted by the inverse of the standard error of an estimate. WLS = weighted least squares. Standard errors in parentheses are robust and clustered at the study and country level (two-way clustering follows Cameron *et al.* 2011). Data are winsorized at 1% level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Gallet (2007) because we have different inclusion criteria. In addition, we have 10 more years of studies in our dataset. The original study uses just a subset of explanatory variables we use. We assume the differences between our study and Gallet (2007) mainly come from the presence of a strong publication bias in this literature (which the original study did not treat for) and possibly some of it is due to omitted variable bias and left-over model uncertainty (for which the original study did not account as well).

Second paper. *(...) I rate the paper as excellent and find it hard to find the paper's weaker points. This is also because it was published by a collective of authors who are at the cutting edge of current meta-analytic research. Nevertheless, I make a few comments below.*

Thank you for your kind words on the paper.

(i.) *While the funnel plot in Figure 3.2 has a seemingly asymmetric shape that indicates publication bias, a large fraction of the estimates appears in some spherical cluster between -1 and 0. It would be interesting to see what studies (with what characteristics) are the remaining observations that complete the overall asymmetric funnel plot. In the extreme case, it could be just a few studies.*

Yes, you are correct to mention that there are two biases in play: one against the positive estimates which we detail in the paper and one against the estimates that are close to -1. But the latter bias is, in fact, very intuitive to observe:

if researchers are trying to tell a story in which skilled and unskilled labor are gross substitutes (or gross complements), they will likely search for estimates that are consistent with that story. As a result, estimates not significantly different from -1 are unfavored and we observe a jump at -1. The jump at -1 is statistically significant, as indicated by the caliper test in Table 3.2, though we fail to find strong evidence that estimates just insignificantly different from -1 are under-reported. Discrimination against inverse elasticities more negative than -1 creates a correlation between estimates and standard errors and is thus captured by meta-regression and nonlinear techniques based on the funnel plot, which generally put weight on the top of the funnel (the most precise estimates).

(ii.) *The regression estimates of publication bias are standard, and the paper includes some of the advanced nonlinear techniques. In particular, regressions that analyze primary studies with IV estimates point to significant publication bias as well as significant effect beyond bias. OLS and natural experiments approaches tend to point only to the existence of publication bias. Nevertheless, this is understandable, given the facts mentioned above. However, I was intrigued to find that fixed effects (FE) models generally only indicate publication bias, especially for the sample that analyzed the primary studies with IV estimates. Is there an explanation for this?*

Study-level fixed effects filter out between-study differences, likely the most important source of endogeneity in this regression. But the identification in the fixed-effect estimator rests on larger studies (reporting more estimates) which could come as less intuitive. In consequence, if a meta-dataset includes very large and very small studies at the same time (such as this one), the estimator may not perform well.

At the same time, the paper is a bit lacking in discussion of why the instruments for IV estimates are constructed as the inverse of the square root of the number of observations. I have come across other approaches in the literature. Is this approach the best for some reason?

We are instrumenting the standard errors of the estimates so we are looking first for the most intuitive instrument that has to be correlated with the standard error. This is exactly the inverse if the square root of the number of observations (the functional comes directly from the definition of the standard error). Any other functional of the number of observations must be an intentional switch; possibly, the meta-analyst identified the obvious functional to be a weak instrument and was fishing for other functional which might result in a stronger (and thus usable) instrument. The method of instrumental variables (MAIVE) is detailed by my co-authors in Irsova *et al.* (2023).

(iii.) *I really liked the section describing the analysis of t-statistics and p-values. I appreciate the inclusion of the latest tests by Elliott* et al. *(2022). In particular,* *the simple graph 3.3 is very informative about the possibility of p-hacking and as such is a great motivation for this whole section of the paper.*

Thank you for your kind words.

(iv.) *The heterogeneity analysis is implemented carefully and includes a number of robustness tests. I appreciate that the authors directly address their approach from the previous section of the paper (publication bias) and control for the different estimation techniques of the primary studies - including the interaction of IV method variable with SE. If I may make a minor comment on this section, some of the factors used do appear to be drivers of heterogeneity, but their occurrence in the literature is quite sparse. This is the case, for example, for cross section studies (Table 3.17 shows, if I understand it correctly, that only in 7.6% of cases were the primary estimates based on crosssections), and similarly for the natural experiment variable (it is =1 only in 6.1% of estimates). Might this sparseness of these characteristics compromise the results a bit? Did the authors choose not to control for characteristics that were even more sparse? For example, in less than 5% of cases? What is the best practice in this area?*

There is no harm in having a variable constituting a fraction of 5% unless there is sufficient between-study heterogeneity in such variable. The choice should follow basic econometric principles: is there enough variation (in meta-analysis this is a between-study variation) in the variable to tell us something useful? In our case, we are happy with the choice of inclusion.

(v.) *A similar question occurred to me while reading the section on implied elasticities. Among other things, the authors' subjective definition of best practice favours dynamic models. However, dynamic models are used in 7.6% of the estimates in the primary studies. I understand that dynamic models are probably better practice than static models according to econometric theory, but cannot the overall implied elasticity for "subjective best practice" in Table 3.5 be challenged by the argument that one of the factors influencing it is based on a small number of observations/studies?*

No, there is no challenge to our results at least for two reasons. First, our model averaging method identifies this variable to be unimportant in the Bayesian sense and we do take care of the model uncertainty. Inclusion of this variable is irrelevant for BMA. Second, it is true that we subjectively opt for dynamic models. This choice is, nevertheless, irrelevant for the best-practice as well: the coefficient of the variable *dynamic model* in Table 3.4 is zero.

Third paper. *The third paper in the thesis is an unpublished solo-authored manuscript that examines the expected returns to higher education in Russia. The study is based on a rich dataset of 13790 individuals between 1994 and 2020. The study uses a diff-in-diff specification, pre/post reform, treatment (higher education attained)/control (no higher education attained).*

(i.) *The paper claims in its introduction (p. 138) that it uses the implementation of USE reform to "helps us to assess whether the unified state exam positively contributed to the student's skills and knowledge, indirectly their earnings and* *salaries". My impression from the paper is that it probably cannot answer such a complex question. While it does show the effects of the reform on wage growth in the treatment (but also control) group, it cannot explain why this growth is occurring. The paper's introduction and motivation focus on the fact that USE reform reallocates knowledge and experience better and better matches the level of students with the level of the university. However, I am concerned that the current results and empirical setting cannot fully answer this.*

As you correctly noted, such goal was too ambitious for my study; therefore, I rewrote my imprecise formulation. Indeed, my study aims to examine the impact of the USE reform on returns to education.

(ii.) *Adding to my confusion, on page 141 the author literally states that "We will assess the impacts of USE reform on higher education enrollment across states and/or federal districts in Russia". However, as I mentioned above, the results of the paper describe the estimated impact of the USE reform on the average monthly salaries (this is exactly how the results section is quoted right under the heading 4.4 Results). In the descriptive statistics in Table 4.1 we can indeed observe numbers that also tell us about enrollments (and other indicators), but this is probably not the main focus of the paper and these numbers are not tested in any statistical/empirical way. I would encourage the author to be consistent in the paper and define exactly what she is doing in the empirical analysis (and stick to that throughout the paper) and what are some side results from descriptive statistics.*

You are correct once again, you have noticed another remnant of my original ambitions that unfortunately cannot be pursued with my data. The text in the fourth chapter has been revised to more accurately reflect the objectives of my analysis and the outcomes obtained.

(iii.) *The overall results of the paper are useful and interesting. They are also complemented by several robustness tests. I find them quite convincing. However, several questions came to my mind. For example, it is interesting (if I am reading the results in Table 4.2 correctly) that earnings after the reform increased by about 5.5% even for people who did not complete higher education (Treatment=0, so I am basing this on the Post-reform coefficient). Is there any way to explain this? Presumably yes, but I would be interested in the author's opinion. What is also interesting, this effect does not occur in Moscow and St. Petersburg.*

Yes, the younger generation earns more than the older generation for several reasons. Firstly, the younger generation is more intensely involved with technology, making them more skilled and valuable in tech-related fields that pay better. Secondly, being in the early stages of their careers, the younger generation typically has more opportunities for advancement and higher earnings. Thirdly, the younger generation entered the workforce during a period of economic growth, which could have come with better job opportunities. Fourthly, the younger generation is more diverse, with a larger representation of women and minorities who have historically been underpaid. I have incorporated your comment into the revised draft.

(iv.) *Next, the reader would need more explanation of the robustness check on page 150. The author reestimated all models using log of average monthly income (note under Table 4.4). The previous results (Table 4.2 and 4.3) were estimated on the log of monthly salary of individuals. However, as I mentioned above, section 4.4 Results is introduced by the sentence "This section shows the estimated impact of the USE reform on the average monthly salaries.". What exactly is the dependent variable in Tables 4.2 and 4.3, and how does it differ from the dependent variable in Table 4.4? In Table 4.4, there are about twice the number of observations (including subsamples). Could the author explain this better? I am sure it can be explained in one sentence so that the reader can understand it accurately. While I understand how to interpret the results in Tables 4.2 and 4.3, I do not know exactly how to interpret the numbers in Table 4.4.*

Great question. The main difference between income and salary in my paper is that the income is just a more general term that encompasses all types of money received by an individual (not a household), while salary specifically refers to a regular payment made to an employee for their work. My variable of total income includes a variety of additional money sources beyond labor such as transfers, yields on investments, rents, or remittances. I have incorporated your comment into my paper to enhance clarity in these definitions.

(v.) *Minor comments include that I could not trace the origin of the statistics in the bottom two paragraphs on page 140. The statistics are quite detailed and interesting (from students' motivations for choosing a university to the level of tuition fees in each area). Do these statistics come from the Russian Longitudinal Monitoring Survey?*

I have rewritten my text as follows: "According to Rosstat (2021), the total number of universities in Russia stood at 1,058 as of 2020, with the majority located in the Moscow federal district (107), St. Petersburg (48), and the Tatarstan Republic (27).Despite the large number of institutions, the government's expenditure on education as a share of gross domestic product (GDP) remains relatively low, at 4%. Every year, approximately 900,000 students are admitted to higher education institutions, with 65% of admissions being funded by the state. The total number of students enrolled in higher education in Russia currently stands at 4 million. In Moscow, the average annual tuition fee at universities is 278,000 Russian rubles (approximately 3,864 USD), making it the most expensive district for higher education in Russia, followed closely behind by St. Petersburg, with an average tuition fee of 208,000 Russian rubles, while Tomsk comes in third place, with 164,000 Russian rubles (Rosstat 2021). These trends are also similar across other federal districts." The statistics used in this paragraph are sourced from the statistical yearbooks published by the Federal State Statistics Service of Russia (Rosstat). I apologize for not properly referencing my sources.

(vi.) *Occasionally, I came across a small typo in the text or a less clear formulation of some sentences. I assume, however, that this will improve when the article undergoes proofreading before publication in the journal - which I believe is a*

> *thing that the author will soon manage to do, and I wish her good luck in this task.*
>
> I agree the submission was written in haste and I have done my best to remedy the typos.

I value immensely the time you took reading my thesis and your kind words on the substance. Thank you for your thorough report and useful comments!

Response to Comments from dr. Nikolai Cook

e) Are there any additional major comments on what should be improved? I have no additional major comments for this dissertation. As a minor comment, in Equation (4.1) the vector of individual, household, and regional characteristics should be denoted as its transpose, X'. As a minor comment, the results writeup on page 147 could be clarified. I find it particularly hard to follow that the treatment group "graduated before the reform".

You are right, I apologize for the inconsistencies and poor phrasing in the original submission, which was only a rough draft. I have since made a concerted effort to enhance the clarity and readability of the text.

In my submission, I define the 'Treatment' group as comprising students who were fully affected by the reform, i.e., those who took the unified state exam (USE) as a final exam in high school and those whose admission to university was determined by their USE scores. Indeed, a self-selection bias could occur if the participants who self-select into the treatment group have different characteristics or motivations than those who self-select into the control group. I admit the assignment of my participants into these groups can be non-random as I use high-school graduates as well as university graduates in one basket. To address the self-selection, I have employed Propensity Score Matching technique which ensures that my treatment and control groups do not differ systematically in their characteristics. Additionally, I have conducted two robustness checks that restrict my sample to address the bias, one of which restricts the sample to university students only, and the other restricts the sample to a specific age cohort.

Also, in response to the comments raised during my pre-defense, I have been thinking about several alternative definitions of control and treatment groups. Indeed, the key is to ensure that the control and treatment groups are comparable in terms of observable and unobservable characteristics that may affect my outcome variables (income and salaries). First, I can take into account different age cohorts. Instead of comparing the outcomes of individuals before and after the implementation of the exam, I can compare the outcomes of individuals who took the exam to those who did not take the exam in the same year (separately for the high school graduates and for the university applicants). Second, I can take into account different regions, specifically, I can compare the outcomes of individuals who took the exam in regions where it was implemented to those in regions where it was not implemented.

Third, I could compare different types of universities (such as public versus private or even less selective versus more selective universities) or different types of high schools (such as vocational versus other high shcools) but unfortunately, these specific data I do not have at disposal. So I am aware there are ways to much improve the paper and I will do my best to do so before I submit this chapter into a journal.

Response to Comments from dr. Jerome Geyer-Klingeberg

Lastly, I should comment on Chapter 4 "Expected Returns to Higher Education in Russia after Unified State Exam Reform". (...) The results chapter and the following discussion are on point, but not as comprehensive as in the previous chapters. However, it is noted on page 2 that this chapter is "to be expanded for the final dissertation". Therefore, I assume that the final thesis will be more elaborated in this regard.

You are too kind; indeed, my introduction and Chapter 4 were written in haste and it showed. In response, I have revised the original text to make it more comprehensive and focused. In addition, following feedback from the public discussion during my pre-defense, I have taken steps to ensure that my analysis in Chapter 4 is more convincing and robust. Specifically, I have used statistical techniques, such as propensity score matching, to adjust for any observed differences in participant characteristics between the treatment and control groups I selected. I have also provided several robustness checks of restricted sub-samples to further support my claims. I have outlined these changes in more detail in my response to the previous reviewers.

Dear referees, dear committee, I have done my best to answer your questions and remedy the mistakes you found. I believe the third essay of the dissertation (Chapter 4) has improved substantially and I thank you greatly for your insightful comments.

References

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

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

CAMERON, A. C., J. B. GELBACH, & D. L. MILLER (2011): "Robust Inference With Multiway Clustering." *Journal of Business & Economic Statistics* **29(2)**: pp. 238–249.

ELLIOTT, G., N. KUDRIN, & K. WUTHRICH (2022): "Detecting p-hacking." *Econometrica* **(forthcoming)**.

FURUKAWA, C. (2020): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Working paper*, MIT.

GALLET, C. (2007): "A Comparative Analysis of the Demand for Higher Education: Results from a Meta-Analysis of Elasticities." *Economics Bulletin* **9(7)**: pp. 1–14.

IOANNIDIS, J. P., T. D. STANLEY, & H. DOUCOULIAGOS (2017): "The Power of Bias in Economics Research." *The Economic Journal* **127(605)**: pp. F236–F265.

IRSOVA, Z., P. R. D. BOM, T. HAVRANEK, & H. RACHINGER (2023): "Spurious Precision in Meta-Analysis." *Technical report*, Available at http://meta-analysis.cz/maive/maive.pdf, Charles University, Prague.

REED, W. R. (2020): "A note on the use of partial correlation coefficients in meta-analyses." *Working Papers in Economics 2020/8*, University of Canterbury, Department of Economics and Finance.

ROSSTAT (2021): "In: Statistical Yearbook of Russia." *Statistical yearbook*, Rosstat, Moscow.

STANLEY, T. D., S. B. JARRELL, & H. DOUCOULIAGOS (2010): "Could It Be Better to Discard 90% of the Data? A Statistical Paradox." *The American Statistician* **64(1)**: pp. 70–77.