

DISCUSSION PAPER SERIES

DP15724

Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought

Tomas Havranek, Zuzana Irsova, Lubica Laslopova
and Olesia Zeynalova

LABOUR ECONOMICS

CEPR

Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought

Tomas Havranek, Zuzana Irsova, Lubica Laslopova and Olesia Zeynalova

Discussion Paper DP15724
Published 26 January 2021
Submitted 20 January 2021

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

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

- Labour Economics

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

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

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

Copyright: Tomas Havranek, Zuzana Irsova, Lubica Laslopova and Olesia Zeynalova

Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought

Abstract

A key parameter in the analysis of wage inequality is the elasticity of substitution between skilled and unskilled labor. We question the common view that the elasticity exceeds 1. Two biases, publication and attenuation, conspire to pull the mean elasticity reported in the literature to 1.9. After correcting for the biases, the literature is consistent with the elasticity in the US of 0.6-0.9. Our analysis relies on 729 estimates of the elasticity collected from 76 studies as well as 37 controls that reflect the context in which the estimates were obtained. We use recently developed nonlinear techniques to correct for publication bias and employ Bayesian and frequentist model averaging to address model uncertainty. Our results suggest that, first, insignificant estimates of the elasticity are underreported. Second, because researchers typically estimate the elasticity's inverse, measurement error exaggerates the elasticity, and we show the exaggeration is substantial. Third, elasticities are systematically larger for developed countries, translog estimation, and methods that ignore endogeneity.

JEL Classification: J23, J24, J31

Keywords: elasticity of substitution, skill premium, meta-analysis, Model uncertainty, Publication bias

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

Zuzana Irsova - zuzana.irsova@ies-prague.org
Charles University, Prague

Lubica Laslopova - l.laslopova@gmail.com
Charles University, Prague

Olesia Zeynalova - olesia.zeynalova@gmail.com
Charles University, Prague

Skilled and Unskilled Labor Are Less Substitutable than Commonly Thought*

Tomas Havranek, Zuzana Irsova, Lubica Laslopova, and Olesia Zeynalova

Charles University, Prague

January 20, 2021

Abstract

A key parameter in the analysis of wage inequality is the elasticity of substitution between skilled and unskilled labor. We question the common view that the elasticity exceeds 1. Two biases, publication and attenuation, conspire to pull the mean elasticity reported in the literature to 1.9. After correcting for the biases, the literature is consistent with the elasticity in the US of 0.6–0.9. Our analysis relies on 729 estimates of the elasticity collected from 76 studies as well as 37 controls that reflect the context in which the estimates were obtained. We use recently developed nonlinear techniques to correct for publication bias and employ Bayesian and frequentist model averaging to address model uncertainty. Our results suggest that, first, insignificant estimates of the elasticity are underreported. Second, because researchers typically estimate the elasticity’s inverse, measurement error exaggerates the elasticity, and we show the exaggeration is substantial. Third, elasticities are systematically larger for developed countries, translog estimation, and methods that ignore endogeneity.

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

JEL Codes: J23, J24, J31

*Corresponding author: Zuzana Irsova, zuzana.irsova@ies-prague.org. Data and code are available in an online appendix at meta-analysis.cz/skill.

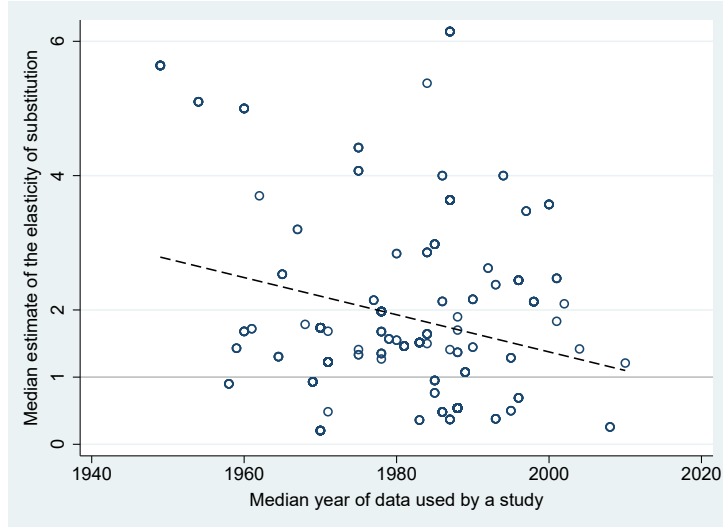
1 Introduction

The elasticity of substitution between skilled and unskilled workers ranks among the most frequently estimated parameters in labor economics: we found 729 estimates reported in 76 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 cross-country heterogeneity in labor productivity (Klenow & Rodriguez-Clare, 1997). Unlike most important parameters in economics, for which often little consensus exists and calibrations vary by the order of magnitude (see, for example, Havranek, 2015, in the context of the elasticity of intertemporal substitution), the elasticity of substitution between skilled and unskilled labor 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 use meta-analysis (Stanley, 2001) to show that the literature is instead consistent with an elasticity below 1.

The observation by Cantore *et al.* (2017) is based on some of the most prominent papers in the literature (Katz & Murphy, 1992; Ciccone & Peri, 2005; Autor *et al.*, 2008), but holds equally true for the literature as a whole as far as the first moment is concerned: the 729 estimates we collect have a mean of 1.9 and median of 1.4. Nevertheless, Figure 1 illustrates that individual studies estimating the elasticity show more disagreement 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; indeed, the figure indicates a slight trend toward smaller elasticities in recent years. So the literature is consistent with a wide range of potential calibrations, though of course the first moment is key in informing them. The problem is that the mean reported estimates in economics are routinely exaggerated by publication bias (Brodeur *et al.*, 2016; Bruns & Ioannidis, 2016; Christensen & Miguel, 2018; DellaVigna *et al.*, 2019; Blanco-Perez & Brodeur, 2020; Brodeur *et al.*, 2020), often by a factor of 2 or more (Ioannidis *et al.*, 2017).¹

¹On the other hand, some recent meta-analyses find no or only mild publication bias: Card *et al.* (2018) in the literature on active labor market programs and Imai *et al.* (2020) in the literature on the present bias.

Figure 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. The horizontal line denotes unitary elasticity. The dashed line denotes a linear trend. Outliers are omitted from the figure for ease of exposition but included in all tests.

Publication bias stems from the tendency of authors, editors, or referees to prefer results that are statistically significant and consistent with the theory (Stanley, 2005). In the context of the elasticity of substitution between skilled and unskilled labor, negative estimates are inconsistent with the canonical model and zero estimates are unintuitive. Few researchers are eager to interpret such estimates, though negative or insignificant estimates of the elasticity will appear from time to time given sufficient imprecision in data and methods. We hypothesize that such estimates are underreported in the literature. The resulting selective reporting creates a paradox: while it most likely improves the inference drawn from an individual study when the researcher does not focus on negative and insignificant estimates, the literature as a whole gets biased towards larger elasticities. The reason is that no upper limit exists that would mirror the lower limit of zero (or statistical insignificance) and ensure symmetry in censoring. McCloskey & Ziliak (2019) liken the problem to the Lombard effect in biology, in which speakers increase their vocal effort in the presence of noise. So, too, can researchers intensify specification searching in response to noise in their data and try a different estimation setup to obtain a larger estimate. Our identification is based on a basic property of the techniques used to estimate the elasticity: the ratio of the estimate to its standard error has a t-distribution, and thus in the absence of

bias the numerator and denominator of the ratio should be statistically independent quantities. Aside from intuitive linear techniques based on the Lombard effect, we employ caliper tests due to Gerber & Malhotra (2008) and new nonlinear techniques due to Ioannidis *et al.* (2017), Andrews & Kasy (2019), and Furukawa (2020).

We have noted that publication bias has been identified in many fields of economics. In most cases, however, it is moderated by attenuation bias in the opposite direction. The “iron law of econometrics” (Hausman, 2001) ensures that virtually all estimates are biased downwards in magnitude because the independent variable is almost always measured with error. The interplay between publication and attenuation biases must be ubiquitous in economics, but to our knowledge has not been explored before. The most famous and frequently used measurements of the elasticity of substitution between skilled and unskilled labor estimate the inverse of the elasticity, thus compounding potential publication bias with inevitable attenuation bias toward larger estimates. For example, extreme measurement error would lead to a zero estimated regression coefficient and thus infinite elasticity. The literature of course realizes the attenuation bias problem, since data on both wages and labor supply can be notoriously noisy, and measurement error is mentioned frequently in the literature (among others 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), though it is hard to find an instrument that would be free of the error. A crude but straightforward way to measure attenuation bias is to compare the implied estimates of the elasticity when the dependent and independent variables are switched. The approach was used in this literature by Bowles (1970) but not since. We employ meta-analysis techniques to exploit the fact that a portion of the literature estimates the elasticity directly.

Our results suggest strong publication and attenuation bias. After correcting for the former, the mean reported elasticity decreases from 1.9 to 1, which is consistent with the typical impact of publication bias in economics identified by Ioannidis *et al.* (2017): twofold exaggeration. The mean of inverted estimates is 1.5 larger than the mean of direct estimates, indicating substantial attenuation bias; the difference remains 1.5 even when we control for publication bias and other 35 controls 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 thoroughly described in Steel (2020). For the former we also employ the dilution prior (George, 2010) that alleviates potential collinearity. Finally, we create a synthetic study that uses all estimates in the literature but assigns more weight to those that are better specified (using Card, 2009, and Autor, 2014, as benchmarks). The resulting estimates for the US range between 0.6 and 0.9, casting doubt on the common view that skilled and unskilled labor are gross substitutes. We also find that larger estimates of the elasticity are associated with developed countries, translog function, higher frequency data, ignoring endogeneity, and controlling for the demand for skills.

The remainder of the paper is structured as follows: Section 2 briefly discusses how the elasticity of substitution between skilled and unskilled labor is estimated, how we collect estimates from primary studies, and presents a bird’s eye view of the literature. Section 3 examines publication bias. Section 4 investigates the sources of heterogeneity in the literature and calculates the mean elasticity implied by the absence of publication bias and identification problems. Section 5 concludes the paper. Appendix A presents additional diagnostics and robustness checks. The data and codes are available in an online appendix at meta-analysis.cz/skill.

2 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

$$\sigma_{US} = \frac{\frac{d(L_U/L_S)}{L_U/L_S}}{\frac{d(w_S/w_U)}{w_S/w_U}} = -\frac{d \log(L_U/L_S)}{d \log(w_U/w_S)}, \quad (1)$$

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 on 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.

To estimate the elasticity researchers usually assume a constant elasticity of substitution (CES) production function:

$$Y = [\alpha(aL_S)^\rho + (1 - \alpha)(bL_U)^\rho]^{\frac{1}{\rho}}, \quad (2)$$

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 easily 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 $\frac{w_S}{w_U}$:

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

Taking logarithms produces a specification that can be estimated:

$$\ln\left(\frac{w_S}{w_U}\right) = \ln\left(\frac{\alpha}{1 - \alpha}\right) + \frac{\sigma - 1}{\sigma} \ln\left(\frac{a}{b}\right) - \frac{1}{\sigma} \ln\left(\frac{L_S}{L_U}\right). \quad (3)$$

The main coefficient of interest, the inverse of the elasticity ($1/\sigma$), can thus be interpreted as the effect of the relative supply of skilled labor on the wage premium to skills. Researchers often include into (3) other variables that capture different characteristics of workers or labor markets, most frequently a proxy for the demand for skills.

Such straightforward specifications have been employed, for example, by Katz & Murphy (1992), Borjas (2003), and Gallego (2012). An important exception are papers along the lines of Krusell *et al.* (2000), who estimate parameters of a structural model that was primarily designed to explore capital-skill complementarity but that is also informative concerning the elasticity of substitution between skilled and unskilled workers. Alternatively, some researchers employ the relative share of skilled labor as the dependent variable and regress it on the wage premium (see, for example Li, 2010). Other researchers use the translog cost function to obtain estimates of the elasticity. This approach is adopted by, for example, Bergstrom & Panas (1992), who simultaneously estimate the cost function and cost share equations by iterative Zellner-efficient procedures. More details on the various estimation techniques employed in the literature are available in Section 4.

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 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 also exploit the previously published narrative reviews by Freeman (1986), Hamermesh (1996), and Behar (2010), and terminate the literature search on March 31, 2019. The data and codes are available in the online appendix 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 729 estimates of the elasticity collected from 76 studies listed in Table 1; we call them primary studies. The oldest study was published in 1970, the most recent one in 2018, covering almost 50 years of research. Almost all collected studies are written in English, with the exception of Jamet (2005), written in French. The histogram of the collected estimates is presented in Figure 2 and shows that the distribution is asymmetric, skewed to the right. The estimated elasticities are widely dispersed, ranging from -437 (Blankenau & Cassou, 2011) to $1,000$ (Psacharopoulos & Hinchliffe, 1972); such extreme values are given by the fact that in most papers the regression estimate has to be inverted to

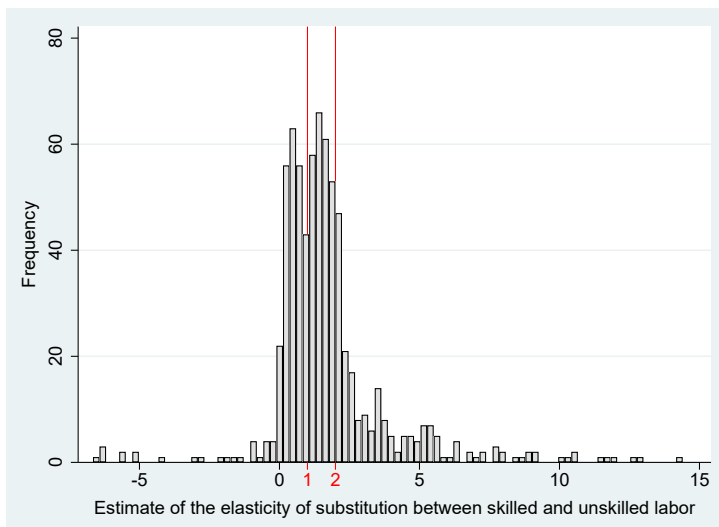
Table 1: Studies used in the meta-analysis

Acemoglu (2002)	Dougherty (1972)	Krusell <i>et al.</i> (2000)
Angrist (1995)	Dupuy (2007)	Kwack (2012)
Askilden & Nilsen (2005)	Dupuy & Marey (2008)	Li (2010)
Autor (2014)	Fallon & Layard (1975)	Lindquist (2005)
Autor <i>et al.</i> (2008)	Fernandez & Messina (2017)	Malmberg (2018)
Avalos & Savvides (2006)	FitzGerald & Kearney (2000)	Manacorda <i>et al.</i> (2010)
Behar (2010)	Foldvari & van Leeuwen (2006)	Medina & Posso (2010)
Bergstrom & Panas (1992)	Freeman (1975)	Mello (2011)
Berndt & Christensen (1974)	Freeman & Medoff (1982)	Mollick (2008)
Berndt & Morrisson (1979)	Gallego (2012)	Murphy <i>et al.</i> (1998)
Binelli (2015)	Gancia <i>et al.</i> (2013)	Nissim (1984)
Blankenau & Cassou (2011)	Giannarakis (2015)	Psacharopoulos & Hinchliffe (1972)
Blundell <i>et al.</i> (2016)	Glitz & Wissmann (2017)	Razzak & Timmins (2008)
Borghans & ter Weel (2008)	Goldin & Katz (2009)	Reshef (2007)
Borjas (2003)	Gyimah-Brempong & Gyapong (1992)	Riano (2009)
Borjas & Katz (2007)	Heckman <i>et al.</i> (1998)	Robbins (1996)
Bound <i>et al.</i> (2004)	Jamet (2005)	Santamaria (2004)
Bowles (1970)	Jensen & Morrissey (1986)	Silva (2008)
Card (2009)	Johnson (1970)	te Velde & Morrissey (2004)
Card & Lemieux (2001)	Katz & Murphy (1992)	Tinbergen (1974)
Choi <i>et al.</i> (2005)	Kawaguchi & Mori (2016)	Verdugo (2014)
Ciccone & Peri (2005)	Kearney (1997)	Wei <i>et al.</i> (2016)
Corker & Bayoumi (1991)	Kesselman <i>et al.</i> (1977)	Welch (1970)
Cruz <i>et al.</i> (2017)	Kim (2005)	Yang (2012)
Das (1999)	Klenow & Rodriguez-Clare (1997)	
Denny & Fuss (1977)	Klotz <i>et al.</i> (1980)	

obtain the elasticity, and when the regression estimate is close to zero (for example, because of measurement error), the implied elasticity is huge in absolute value. To prevent these outliers from driving our results, we winsorize the sample at 4% (the level at which our results stabilize and hold irrespective of further winsorization) and work with the winsorized sample from now on. After winsorization the reported elasticities range from -0.34 to 8.81 and are characterized by a mean of 1.93 and a median of 1.41 . It is worth noting that the median estimate (by definition not affected by winsorization) is identical to the estimate of Katz & Murphy (1992), the most iconic and cited result in the literature, and also close to the consensus value of 1.5 presciently identified by Cantore *et al.* (2017).

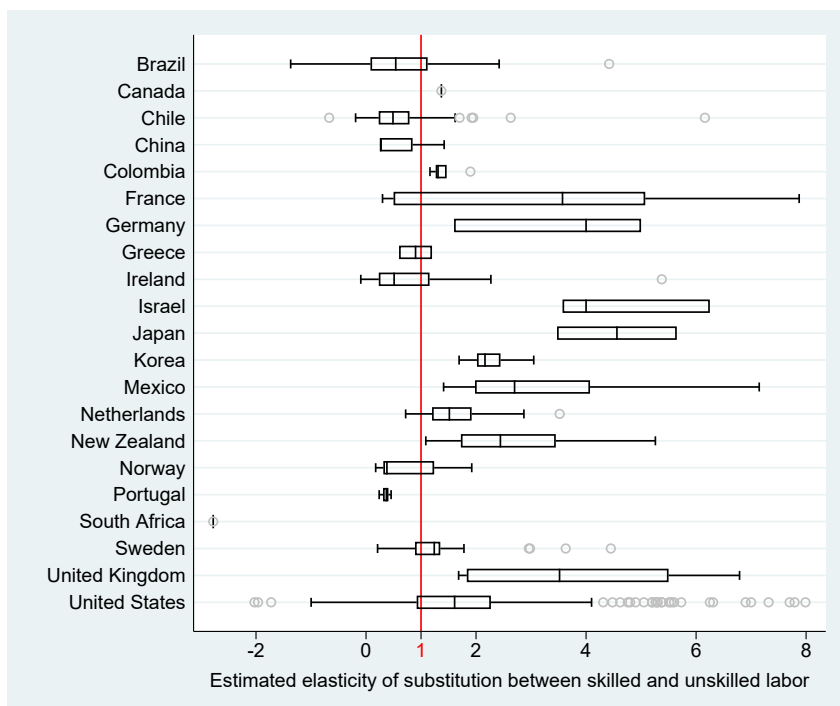
Of the 729 estimates we collect, merely 253 (or 34.7%) fall into the $(1, 2)$ interval, from which most of the calibrations in the literature are drawn. Next, 471 (64.6%) estimates in our sample support the hypothesis that skilled and unskilled labor are gross substitutes. In contrast, 216 (29.6%) estimates fall into the $(0, 1)$ interval, suggesting gross complementarity between the two types of labor. While many estimates are reported that are in the vicinity of 0.5 and 1.5 , Figure 2 shows a surprising drop in the frequency of estimates around 1 , the case

Figure 2: Distribution of the reported estimates



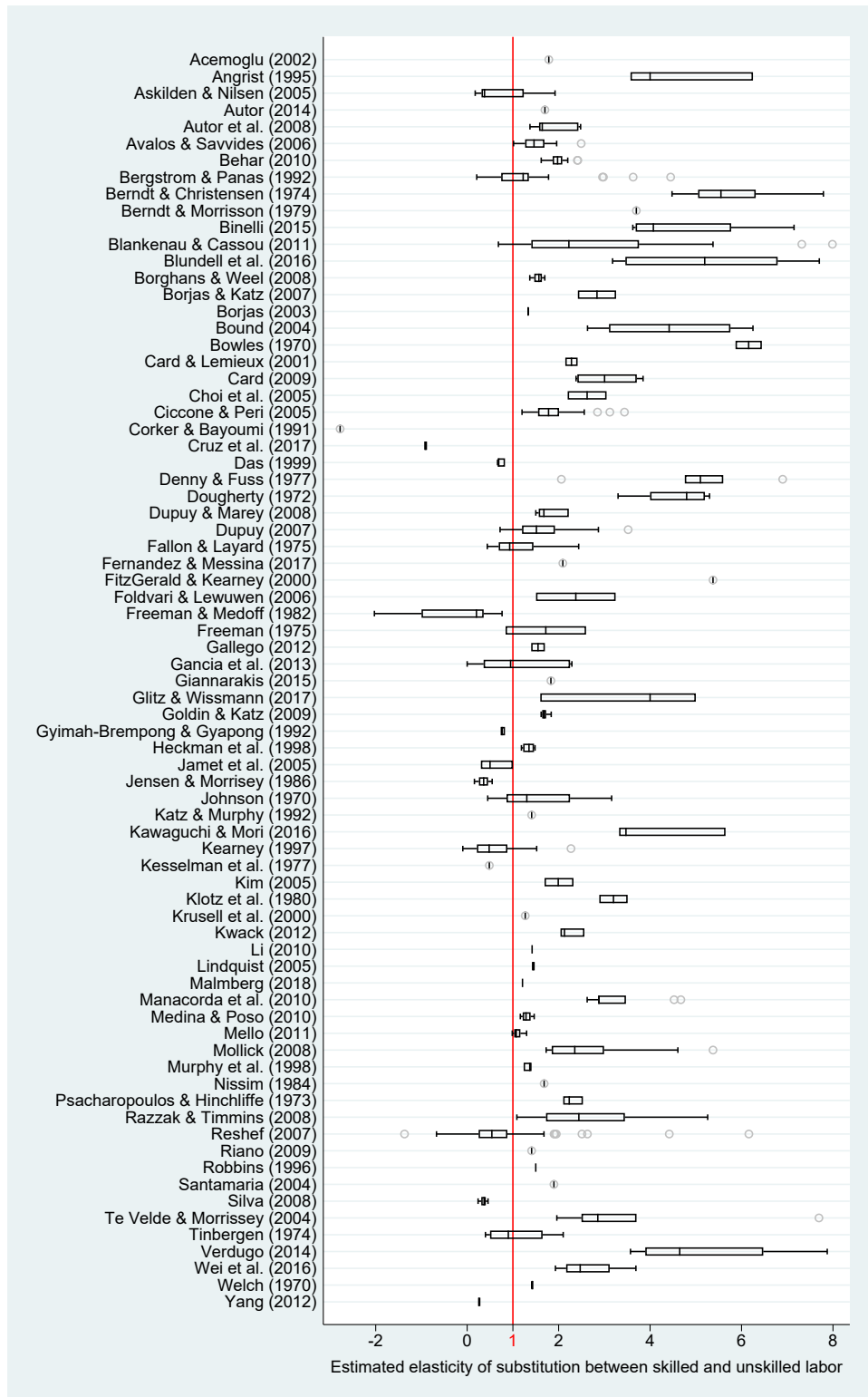
Notes: The figure depicts a histogram of the elasticities reported by individual studies. The vertical lines denote the interval $(1, 2)$, from which most of the values used for calibrations are drawn.

Figure 3: Cross-country heterogeneity in the 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. Outliers are excluded from the figure. The vertical line denotes unitary elasticity. For ease of exposition, outliers are excluded from the figure but included in all statistical tests.

Figure 4: Estimates of the elasticity vary both within and across studies



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

where the CES function reduces to Cobb-Douglas. In any event, the figure makes it clear that there is substantial heterogeneity in the literature. Figure 4 shows that results vary significantly not only across studies, but also within studies, and the same observation holds for individual countries (Figure 3).

Before proceeding to a detailed analysis, let us take a first look at the potential causes of heterogeneity. Table 2 provides summary statistics for various subsamples of the data. It also provides a weighted mean: weighting by the inverse of the number of estimates reported per study ensures that all studies get the same weight. The exact definitions of the listed groups can be found later in Table 5. First, Table 2 suggests that short-run estimates, ones produced using first-difference estimation or the error-correction model, do not differ much on average from long-run estimates. The Hicks definition of the elasticity seems to yield larger estimates than the remaining definitions in our sample. The smaller mean corresponding to the remaining definitions is, however, not driven by Allen-Uzawa elasticities (which, in theory, can be negative) but by the estimates of Morishima and Shadow elasticities. Moreover, most of the collected estimates are derived as the inverse of the estimated regression coefficient, and this seems to be associated with systematically larger elasticities compared to the estimates derived directly from the estimated regression. On average, the difference between direct and inverted estimates is 1.5, a finding that indicates substantial attenuation bias and suggests (1.10, 2.65) as the interval for the underlying mean elasticity corrected for the bias.

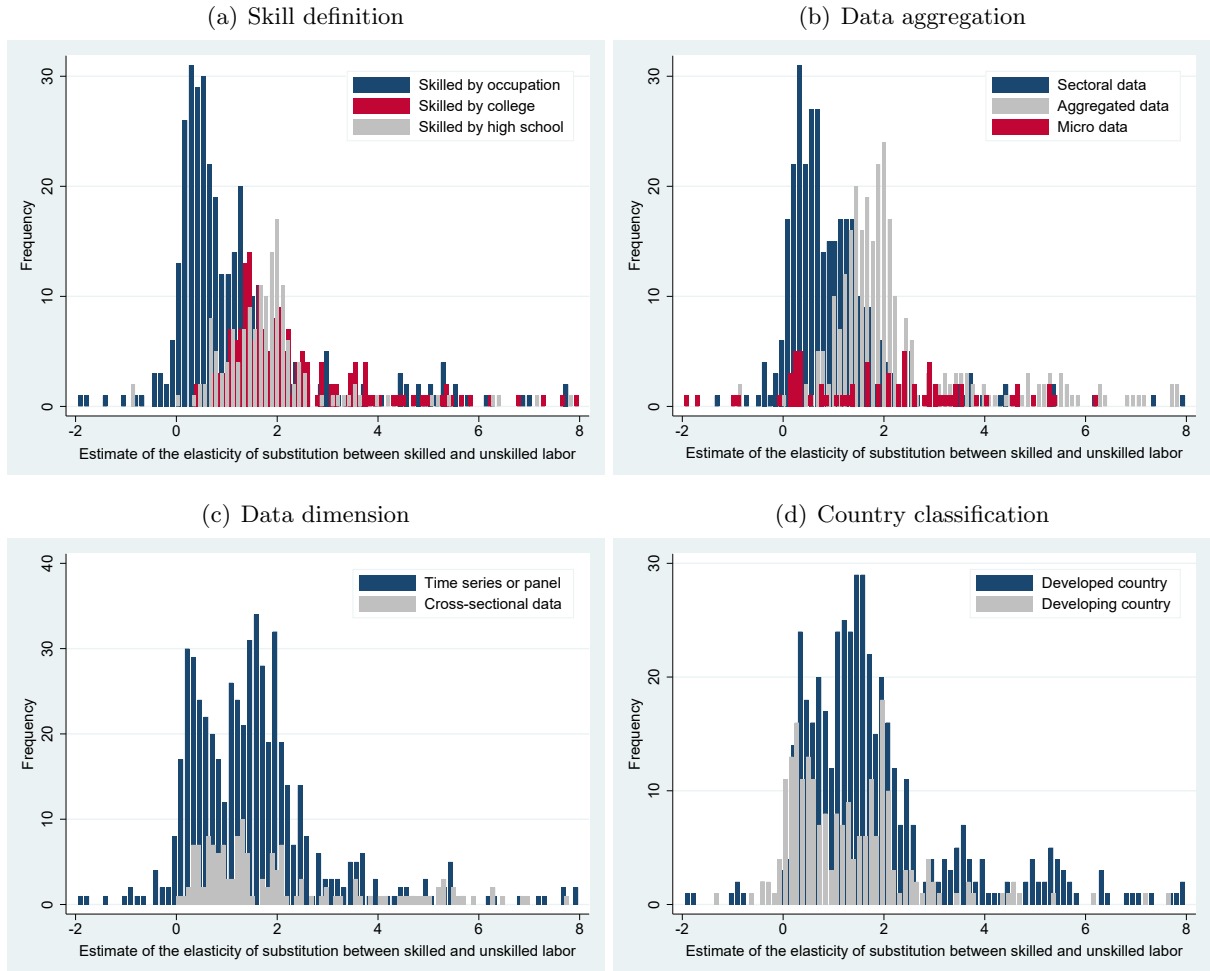
Second, the table suggests that the definition of “skills” matters for the estimated elasticity. Once the skill of a worker is defined by experience, profession, or training rather than by education background, skilled and unskilled labor become less substitutable (a pattern also apparent visually in Figure 5a). In contrast, the elasticity does not change much when high school instead of college is taken as the threshold for skill (typically in developing countries). Third, the elasticity differs with different data characteristics. For example, substitutability increases with an increase in data frequency. Substitutability also seems to be a little higher when the time dimension of the data is not accounted for (Figure 5c). Both weighted and unweighted means indicate that with with sectoral data researchers tend to find less evidence for substitutability compared to the case when micro data are used (Figure 5b), although one would expect more substitutability with more aggregated data.

Table 2: Summary statistics for different subsets of the literature

	No. of observations	Unweighted			Weighted		
		Mean	95% conf. int.		Mean	95% conf. int.	
<i>Horizon</i>							
Short-run effect	55	1.89	1.44	2.35	2.24	1.72	2.76
Long-run effect	674	1.93	1.77	2.09	2.37	2.20	2.53
<i>Variable definition</i>							
Hicks elasticity	641	1.96	1.79	2.12	2.45	2.28	2.62
Other elasticity	88	1.72	1.33	2.12	1.82	1.44	2.20
Skilled by college	199	2.91	2.56	3.25	2.35	2.10	2.60
Skilled by high school	172	2.18	1.89	2.48	3.23	2.79	3.67
Skilled by occupation	358	1.26	1.10	1.43	1.80	1.61	2.00
<i>Data characteristics</i>							
Higher frequency	53	3.48	2.73	4.22	3.27	2.59	3.94
Annual frequency	580	1.86	1.69	2.03	2.35	2.17	2.53
Lower frequency	96	1.46	1.30	1.63	1.39	1.25	1.54
Micro data	76	2.21	1.72	2.70	2.02	1.57	2.47
Sectoral data	342	1.31	1.12	1.51	1.72	1.54	1.90
Aggregated data	311	2.54	2.30	2.77	2.63	2.38	2.88
Cross-section	151	2.42	2.05	2.79	2.61	2.21	3.01
Panel or time-series	578	1.80	1.64	1.96	2.27	2.10	2.43
<i>Structural variation</i>							
United States	287	2.25	1.99	2.51	2.30	2.08	2.52
Developing country	227	1.36	1.18	1.55	2.21	2.01	2.40
Developed country	502	2.18	1.99	2.38	2.39	2.19	2.59
Male workers	143	2.46	2.11	2.80	2.62	2.27	2.98
Male & female workers	586	1.80	1.63	1.97	2.27	2.09	2.44
Manufacturing sector	275	1.15	0.97	1.34	2.34	2.08	2.59
Other than manufacturing sector	454	2.40	2.20	2.60	2.36	2.16	2.56
<i>Design of the production function</i>							
One-level CES function	419	1.51	1.35	1.67	2.08	1.91	2.24
Multi-level CES function	190	2.53	2.18	2.88	2.25	1.94	2.57
Other function	120	2.43	2.00	2.87	2.85	2.41	3.29
<i>Estimation technique</i>							
Dynamic model	124	1.36	1.07	1.65	1.81	1.54	2.08
Unit fixed effects	372	1.70	1.52	1.88	2.08	1.89	2.27
Time fixed effects	116	2.38	2.02	2.73	2.73	2.25	3.20
OLS method	447	1.66	1.48	1.85	2.20	2.01	2.39
IV method	178	2.61	2.26	2.97	2.85	2.52	3.18
SUR method	43	1.37	1.03	1.72	1.61	1.15	2.06
ML method	49	2.31	1.89	2.72	2.74	2.24	3.23
Inverted estimate	389	2.65	2.44	2.87	2.55	2.35	2.76
Direct estimate	340	1.10	0.93	1.26	1.99	1.74	2.23
<i>Publication characteristics</i>							
Unpublished study	298	1.18	1.00	1.36	1.74	1.54	1.93
Published study	431	2.45	2.24	2.65	2.63	2.41	2.84
Top journal publication	206	2.16	1.88	2.45	2.61	2.28	2.95
All estimates	729	1.93	1.78	2.08	2.36	2.20	2.51

Notes: The table reports summary statistics of the reported elasticity of substitution for different subsets of the literature. The exact definition of the variables is available in Table 5. Weighted = estimates are weighted by the inverse of the number of estimates reported per study.

Figure 5: Prima facie patterns in the data



Notes: The figure depicts histograms of the reported elasticities divided based on the definition of skills, aggregation of data, dimension of data, and country development status. We use the IMF definition to classify countries as developed or developing.

Fourth, a systematic role of structural variation (underlying differences in the elasticity across countries, industries, and genders) is not clear-cut. The differences in the means of the elasticity between developing and developed countries (Figure 5d) as well as between manufacturing and other sectors almost disappear once each study is given the same weight. If a more homogenous sample is considered, such as that with male workers only, the estimated substitutability gets a bit higher on average. The design of the production function and the estimation technique involved, in contrast, are associated with different elasticities. When other than CES function is used (typically the translog function), the estimated elasticities tend to be larger. Moreover, OLS, dynamic models, and SUR produce smaller estimates in comparison to,

for example, time fixed effects and IV. But inference based on Table 2 may suffer from omitted variable bias, and a more nuanced analysis is needed. We provide such an analysis in Section 4.

To illustrate the potential differences in quality not captured by differences by the variables introduced above, we single out unpublished studies (working papers or published analyses that did not undergo the peer-review process), which seem to produce considerably smaller estimates than peer-reviewed studies. The finding is consistent with higher quality studies reporting larger estimates, but can also be consistent with publication bias stemming from the preferences of editors and referees for such estimates. In this context it is important to note that there is little reason to expect publication bias to only affect published studies. Researchers write their papers with the intention to publish, and rational authors will accommodate the expected preferences of editors and referees already in the working paper version. Even more plausibly, publication bias stems simply from the desire to report results that are reasonable (positive) and interesting (statistically significant), and the publication process plays a small role in such selective reporting. The next section analyzes publication bias in detail.

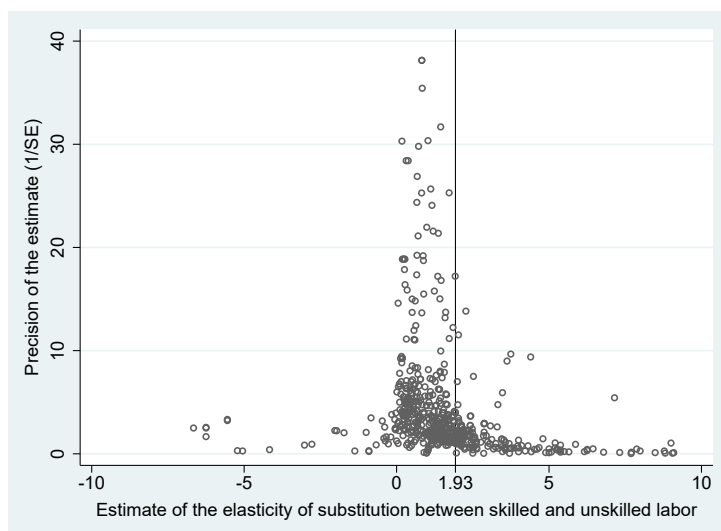
3 Publication Bias

An intuitive feature of the elasticity of substitution between skilled and unskilled labor is its nonnegativity. As Kearney (1997, p. 33) notes on his negative estimates: “The implied coefficients . . . violate standard economic theory.” Some researchers, such as Bowles (1970, p. 73), act upon such fact and “exclude [negative elasticity] values . . . as implausible on a priori grounds.” While a zero elasticity is possible in theory, it is also often deemed implausible and rarely interpreted. What follows is a general tendency in the literature to discriminate against negative and insignificant (small positive) values of the elasticity. While large positive estimates are sometimes also deemed implausible, and already Tinbergen (1974, p. 217) was “struck by the high elasticity figures obtained by several others,” there exists no upper threshold that would immediately identify implausibility, and large estimates can thus be often retained. Hence the mean and median estimates in the literature are probably biased upwards. Note that such publication bias is natural, inevitable in economics, 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 so far attempted to do so in the case of the elasticity of substitution between skilled and unskilled labor.

We use several methods to identify and correct for publication bias, but the underlying identification strategy remains the same: in the absence of bias, there should be no relationship between estimates and their standard errors. Departures from the absence of a relationship (linear or nonlinear) identify the magnitude of bias. The strategy follows from the fact that the ratio of the estimates to their standard errors has a t-distribution, which means that the nominator and denominator form statistically independent quantities. If, in contrast, positive and statistically significant estimates are preferred, researchers will compensate large standard errors (given by noise in the data or estimation technique) by searching for large point estimates. This specification search can involve trying different subsamples, different treatment of outliers, different techniques, or different control variables. Again we note that no ulterior motives are necessary: negative and insignificant estimates can simply be viewed as evidence that there is something wrong with the estimated specification and that it should be adjusted.

Figure 6: The funnel plot suggests publication bias



Notes: When there is no publication bias, estimates should be symmetrically distributed around the mean (denoted by the vertical line). Outliers are excluded from the figure for ease of exposition but included in all statistical tests.

It is certainly possible that our identification assumption, quite commonly used in the meta-analysis literature, can be violated in labor economics. If a research technique influences both the estimate and the corresponding standard error in the same direction, a positive correlation

between the two quantities arises but cannot be interpreted as publication bias. For example, consider IV and OLS estimates in the literature. IV estimates come with larger standard errors, and if they also correct for downward endogeneity bias, the entire literature will feature a positive correlation between estimates and standard errors even in the absence of publication bias. In fact we find (in the next section) that IV estimates of the elasticity tend to be *smaller* than OLS estimates, which means that our measures of publication bias are probably understated. But estimates and standard errors can be affected by many method choices, and one cannot hope to explicitly control for all of them. A natural solution is to instrument the standard error using the inverse of the square root of the number of degrees of freedom used in the primary studies, which is proportional to the standard error by definition and should not be much correlated with method choices. Alas, in our case the instrument is weak. We thus additionally use caliper tests (Gerber & Malhotra, 2008), which are not based on the aforementioned assumption and compare the ratio of estimates in a narrow band above and below a specified threshold for t-statistics (1.96 and 0).

In the remainder of the section we go through individual tests for publication bias. A simple visual tool is the so-called funnel plot (Egger *et al.*, 1997), a scatter plot of the estimated elasticity values on the horizontal axis plotted against their estimated precision (the inverse of the standard error) on the vertical axis. If there is no publication selection bias in the literature, the funnel plot should be symmetrical around the mean value. While imprecise estimates at the bottom of the funnel plot are dispersed, with increasing precision the estimates get closer to the underlying mean. Thus the plot should attain an inverted funnel shape. The funnel plot for our dataset is shown in Figure 6, and it is apparently asymmetric: negative and small positive estimates are underrepresented in the literature. Funnel asymmetry can simply be tested formally by regressing the horizontal axis on the (inverted) vertical axis (Card & Krueger, 1995):

$$\sigma_{ij} = \sigma_0 + \beta \cdot SE(\sigma_{ij}) + \epsilon_{ij}, \quad (4)$$

where σ_{ij} stands for the i -th estimate of the elasticity of substitution reported in the j -th study, $SE(\sigma_{ij})$ denotes its standard error, and ϵ_{ij} is the error term. If the estimated β is significantly different from zero, we obtain evidence of funnel asymmetry and hence publication bias. If we assume that publication bias is a linearly increasing function of the standard error (in the

Table 3: Linear and nonlinear techniques detect publication bias

<i>Panel A: unweighted</i>	OLS	Fixed effects	Hierarchical Bayes
Standard error (<i>publication bias</i>)	0.375 ^{***} (0.0450) [0.220, 0.454]	0.316 ^{***} (0.0375) –	0.540 ^{***} (0.0041) –
Constant (<i>effect beyond bias</i>)	1.350 ^{***} (0.295) [0.695, 2.052]	1.439 ^{***} (0.0568) –	1.360 ^{***} (0.0039) –
Observations	638	638	638
<i>Panel B: weighted</i>	Study-weighted	Precision-weighted	Top journals
Standard error (<i>publication bias</i>)	0.257 ^{***} (0.0885) [0.061, 0.498]	0.827 ^{***} (0.146) [0.500, 1.198]	0.388 ^{***} (0.0929) [-0.123, 0.927]
Constant (<i>effect beyond bias</i>)	1.829 ^{***} (0.224) [1.355, 2.304]	1.086 ^{**} (0.468) [-0.072, 2.170]	1.571 ^{***} (0.281) [0.788, 3.258]
Observations	638	638	152
<i>Panel C: non-linear</i>	WAAP (Ioannidis <i>et al.</i> , 2017)	Stem-based method (Furukawa, 2020)	Selection model (Andrews & Kasy, 2019)
Effect beyond bias	0.950 ^{***} (0.052)	0.833 ^{***} (0.107)	1.020 ^{***} (0.058)
Observations	638	638	638

Notes: The first two panels report the results of the regression $\sigma_{ij} = \sigma_0 + \beta \cdot SE(\sigma_{ij}) + \epsilon_{ij}$, where σ_{ij} denotes the i -th effect estimated in the j -th study, and $SE(\sigma_{ij})$ denotes its standard error. In the second column of Panel A study-level fixed effects are included. In the first column of Panel B the model is weighted by the inverse of the number of estimates per study. Precision-weighting indicates that the model is weighted by the inverse of the standard error of an estimate. The “top journals” subsample includes the most prestigious outlets in our sample: American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Econometrica, Review of Economics and Statistics, and Science. In Panel C, WAAP stands for Weighted Average of the Adequately Powered. Standard errors, clustered at the study level, are reported in parentheses (except for Hierarchical Bayes, which has posterior standard deviation in parentheses; stars for Hierarchical Bayes are presented only as an indication of the parameter’s importance to keep visual consistency with the rest of the table). 95% confidence intervals from wild bootstrap in square brackets (Roodman *et al.*, 2018). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

nonlinear and caliper tests we will abandon this strong assumption), σ_0 measures the mean elasticity corrected for the bias (Stanley, 2005). We cluster standard errors at the study level and also report confidence intervals resulting from wild bootstrap (Roodman *et al.*, 2018).

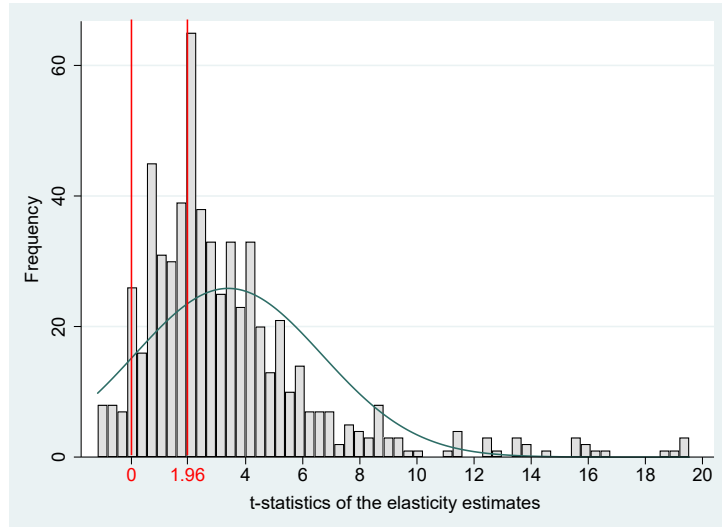
Table 3 presents the results of various estimation techniques used to run Equation 4 on the sample of observations for which standard errors are available (which means that we have to drop the 91 estimates for which no precision was reported; we also note that our results do not change qualitatively if we focus solely on long-run estimates, see Table A1). The baseline model in the first column of Panel A uses OLS. In the second column we control for unobserved heterogeneity across studies by employing study-level fixed effects; in the third column we use a multilevel estimation technique which implements partial pooling at the study level and uses within-study variation to influence pooling weights. This so-called Hierarchical Bayes model

implements the Gibbs sampler with a standard prior according to Rossi *et al.* (2005). The results are consistent with our interpretation of the funnel plot: we find publication bias in the literature. The mean elasticity beyond publication bias in Panel A corresponds to the value of 1.4 echoed by many studies (for example, Katz & Murphy, 1992).

In panel B of Table 3 we apply two different weighting schemes. First, we weight (4) by the inverse of the number of estimates reported per study. This way all studies get the same weight in our analysis irrespective of how many estimates they report. Second, we weight the regression by the inverse of standard error of the estimate (i.e., precision). This way we explicitly address heteroskedasticity inherent in (4). In the third column of Panel B we only include estimates from studies published in the most prestigious journals in our sample: *The American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, *Review of Economics and Statistics*, and *Science*. The results corroborate the presence of publication bias. We find it interesting to observe that the subsample of estimates published in top journals produces results very similar to the baseline case. The problem with both Panel A and Panel B is that these techniques assume a linear relationship between the standard error and publication bias. While a good intuitive start, there is little reason to believe the first-order approximation is realistic.

Panel C of Table 3 features more realistic (nonlinear) techniques that were recently developed to match patterns in large samples of meta-analyses and replication efforts. We apply three different methods: the weighted average of adequately powered estimates by Ioannidis *et al.* (2017), the stem-based method by Furukawa (2020), and the selection model by Andrews & Kasy (2019). Ioannidis *et al.* (2017) propose a correction procedure that focuses on estimates with statistical power above 80%. The method of Furukawa (2020) extends the approach of Stanley *et al.* (2010), who suggest to use 10% of the most precise estimates, and selects the optimal ratio of the most precise studies endogenously by exploiting the trade-off between bias and efficiency. The estimator of Andrews & Kasy (2019) represents the latest incarnation of selection models in the tradition of Hedges (1992); the technique gives more weight to estimates that appear underreported in the literature (such as those insignificant at the 5% level). Several Monte Carlo simulations have shown that nonlinear techniques perform better than linear approximations (Ioannidis *et al.*, 2017; Furukawa, 2020). The three techniques yield remarkably consistent

Figure 7: The distribution of t-statistics peaks at 2



Notes: The figure represents the distribution of t-statistics of the reported estimates of the elasticity overlaid on a corresponding normal distribution. Red lines represents critical value of 1.96 associated with significance at the 5% level and the value of 0 associated with changing the sign of the estimate. We exclude estimates with large t-statistics from the figure for ease of exposition but include them in statistical tests.

results and suggest an underlying mean elasticity of about 1, compared to the naive mean of 1.9 prior to the correction for publication bias.

We have noted that all the major tests of publication bias are based on the assumption that in the absence of bias there is no relationship between estimates and their standard errors. A method that does not rely on this assumption was introduced by Gerber & Malhotra (2008) and is called “caliper test;” unfortunately, it can only examine the presence of bias and does not offer an estimate of the corrected elasticity. The test is based on the idea that publication bias is consistent with sudden jumps in the frequency of the t-statistic, especially around the critical value of 1.96 (recomputing t-statistics to z-statistics would not change our results because in the vast majority of studies both are virtually identical) but also at 0. Indeed, Figure 7 shows that both thresholds are associated with jumps, with a stronger jump at 1.96. Caliper tests seek to identify the significance of the jump by comparing in a narrow band the frequency of estimates above and below the threshold. With a sufficiently narrow caliper both frequencies should be the same. Our results in Table 4 confirm that the frequency jumps are significant both in statistical and economic terms. Near the thresholds, estimates significant at the 5% level dominate the insignificant ones 59% to 41%; positive estimates dominate negative ones

Table 4: Caliper tests indicate publication bias

Threshold for t-statistic: 1.96			
Share of estimates above caliper minus 0.5	caliper size: 0.4 0.0902 ^{**} (0.0447)	caliper size: 0.5 0.0804 [*] (0.0414)	caliper size: 0.6 0.0886 ^{**} (0.0393)
Observations	122	143	158
Threshold for t-statistic: 0			
Share of estimates above caliper minus 0.5	caliper size: 0.6 0.153 ^{***} (0.0553)	caliper size: 0.7 0.167 ^{***} (0.0517)	caliper size: 0.8 0.177 ^{***} (0.0480)
Observations	75	84	96

Notes: The table reports results for caliper tests introduced by Gerber & Malhotra (2008). The tests compare the relative frequency of estimates above and below an important threshold for the t-statistic. We use calipers of different sizes depending on the number of observations available (for the second threshold narrower calipers are unfeasible due to the lack of observations in the immediate vicinity of zero). A test statistic of 0.153, for example, means that 65.3% estimates are above the threshold and 34.7% estimates are below the threshold. Standard errors are reported in parentheses and clustered at the study level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

65% to 35%. We conclude that publication bias in the literature is substantial and caused by preference for both positive and statistically significant estimates.

4 Heterogeneity

The empirical literature on the elasticity of substitution between skilled and unskilled labor is characterized by significant variation in the reported estimates, as we have shown earlier in Figure 3 and Figure 4. While publication bias explains a piece 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, in addition to the standard error, for 36 variables that capture the different 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 relationship between estimates and standard errors, which serves as an indication of publication bias, is robust to controlling for the various aspects of study design. 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 of our analysis we create a synthetic study that computes an implied mean elasticity using all estimates but giving more weight to those that are arguably better identified, all the while correcting for publication and attenuation bias.

4.1 Variables

The studies estimating the elasticity differ in so many dimensions that it is unfeasible to control for all differences. In the previous 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 42 main characteristics (and consequently, to avoid the dummy trap, codify 36 explanatory variables to be used in model averaging) which we distribute for ease of exposition into six categories: variable definition, data characteristics, structural variation, design of the production function, estimation technique, and publication characteristics. Table 5 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 5: Description and summary statistics of regression variables

Variable	Description	Mean	SD	WM
Elasticity estimate	Estimate of the elasticity of substitution between the skilled and unskilled labor (response variable).	1.93	2.07	2.47
Standard error (SE)	Standard error of the estimate of the elasticity of substitution. The variable is important for gauging publication bias.	1.52	3.46	1.76
<i>Variable definition</i>				
Hicks elasticity	=1 if Hicks (1963) elasticity of substitution is estimated.	0.88	0.33	0.85
Other elasticity	=1 if Allen-Uzawa (Allen & Hicks, 1934; Uzawa, 1962), Morishima (1967), or Shadow (Berndt & Christensen, 1974) elasticity of substitution is estimated (reference category for the different types of the elasticity).	0.12	0.33	0.15
Skilled by college	=1 if tertiary education (college) is used in the estimation as the definition of skill.	0.27	0.45	0.47
Skilled by high school	=1 if secondary education (high school) is used as the definition of skill (reference category for the definition of skills).	0.24	0.42	0.21
Skilled by occupation	=1 if occupation type, e.g. white collar, is used for the definition of skill.	0.49	0.50	0.32
<i>Data characteristics</i>				
Higher frequency	=1 if higher than annual frequency of the data is used; typically monthly, quarterly, or semi-annual.	0.07	0.26	0.09
Annual frequency	=1 if annual frequency of the data is used in the estimation (reference category for data frequency).	0.80	0.40	0.83

Continued on next page

Table 5: 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.	0.13	0.34	0.08
Micro data	=1 if micro-level data (unit = single worker or firm) are used in the estimation.	0.10	0.31	0.17
Sectoral data	=1 if sector-level data (unit = sector) are used in the estimation.	0.47	0.50	0.18
Aggregated data	=1 if aggregated data (unit = economy or regions) are used in the estimation (reference category for the type of data aggregation).	0.43	0.49	0.65
Cross-section	=1 if cross-sectional data are used; =0 if time-series or panel data are used.	0.21	0.41	0.27
Data midyear	The logarithm of the median year of the time period of the data used to estimate the elasticity.	3.94	0.34	3.94
Data length	The logarithm of the number of years of the data period used in the estimation.	2.53	1.35	2.31
<i>Structural variation</i>				
United States	=1 if the country for which the elasticity is estimated is the United States.	0.39	0.49	0.41
Developing country	=1 if a developing country is considered, =0 if a developed country is considered in the estimation.	0.31	0.46	0.18
Male workers	=1 if male workers only are considered, =0 if female or a mixed sample is considered in the estimation.	0.20	0.40	0.25
Manufacturing sector	=1 if the elasticity is estimated for the manufacturing sector, =0 if another sector is considered.	0.38	0.49	0.13
<i>Design of the production function</i>				
One-level CES function	=1 if one-level CES functional form of the production function is used in the estimation.	0.57	0.49	0.33
Multi-level CES function	=1 if multi-level CES functional form of the production function is used in the estimation.	0.26	0.44	0.39
Other function	=1 if other functional forms of the production function including the translog form of the production or cost function and undisclosed production form is used (reference category for the functional form).	0.16	0.37	0.28
Time control	=1 if time control is included in the model (capturing, e.g., technological change).	0.52	0.50	0.48
Location control	=1 if location/unit control is included (capturing spatial variation).	0.13	0.34	0.15
Education control	=1 if control for the level of education is included (or the level of experience or part-time status).	0.03	0.16	0.09
Macro control	=1 if macroeconomic indicators are included.	0.09	0.29	0.09
Population control	=1 if socioeconomic variables are included.	0.01	0.11	0.02
Sectoral control	=1 if the model controls for different sectors.	0.04	0.21	0.08
Age control	=1 if a control for the age of workers is included.	0.07	0.25	0.12
Ethnicity control	=1 if control for ethnicity, nationality, or other immigration-related variable is included.	0.03	0.16	0.09
Capital control	=1 if a capital-related control is included (capturing changes in capital stock under a capital-skill complementarity technology).	0.08	0.27	0.11
<i>Estimation technique</i>				
Dynamic model	=1 if the model form used for estimation is dynamic (VAR, ECM, VECM, PAD, ADL, DLTM, DOLS)	0.17	0.38	0.19
Unit fixed effects	=1 if the model is estimated in first differences or cross-sectional fixed effects are considered.	0.51	0.50	0.45
Time fixed effects	=1 if time fixed effects are included in the model.	0.16	0.37	0.12

Continued on next page

Table 5: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD	WM
Long-run effect	=1 if long-run effect is estimated.	0.92	0.26	0.92
OLS method	=1 if the ordinary least squares method or its variations (WLS, GLS) are used.	0.61	0.49	0.58
IV method	=1 if instrumental estimation method is used, including 2SLS, 3SLS, and GMM.	0.24	0.43	0.22
SUR method	=1 if the seemingly unrelated regression method is used.	0.06	0.24	0.05
ML method	=1 if maximum likelihood or its variations (FIML, LIML, SPML) are used for estimation (reference category for the method variables).	0.07	0.25	0.06
Inverted estimate	=1 if the elasticity is computed by inverting the estimated regression coefficient. The variable is important for gauging attenuation bias.	0.53	0.50	0.66
Direct estimate	=1 if the elasticity is not an inversion of the estimated regression coefficient (reference category).	0.47	0.50	0.34
<i>Publication characteristics</i>				
Impact factor	The discounted recursive RePEc impact factor of the outlet.	0.69	1.03	1.03
Citations	The logarithm of the number of per-year citations of the study in Google Scholar.	1.36	1.13	1.77
Published study	=1 if the study was published in a peer-reviewed journal.	0.59	0.49	0.70

Notes: 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, LIML = limited information maximum likelihood, FIML = Fuller LIML, SPML = simulated pseudo-maximum likelihood.

Variable definition A fundamental issue in the estimation of the elasticity is its definition. The *Hicks elasticity*, as referred to in Equation 1 in Section 2, is by far the most common measure in our sample, accounting for about 90% of the collected estimates (and about 97% of the estimates for which standard errors are available). The *Allen-Uzawa elasticity* of substitution is an alternative derived from the cost function and has been applied by Jensen & Morrisey (1986), Bergstrom & Panas (1992), and Askilden & Nilsen (2005), among others. An important distinction is that the Allen-Uzawa elasticity can, in theory, be negative; also it does not measure the changes of relative quantities moving along an isoquant of production, as the Hicksian measure does. Instead, Allen-Uzawa is a one-factor-one-price elasticity. The *Morishima elasticity* (Gyimah-Brempong & Gyapong, 1992), on the other hand, is a two-factor-one-price elasticity: tracking a change in a factor ratio in response to a change in the price of one factor. The Morishima elasticity is more similar to Hicks elasticity and cannot plausibly be negative.

The final special case of the elasticity is the so-called *Shadow elasticity* (Berndt & Christensen, 1974), which allows prices to change such that the average cost is fixed.

Another key issue in the estimation of the elasticity is the definition and measurement of skills. Most commonly, researchers use a cut-off level of a pre-specified education achievement. For instance, Dupuy & Marey (2008) estimate the elasticity between college-educated workers and other workers (variable *Skilled by college*), Mello (2011) uses primary education for the cut-off, and Ciccone & Peri (2005) consider all workers with at least secondary education as skilled (*Skilled by high school*). Next, some researchers identify skilled and unskilled workers based on the type of occupation (*Skilled by occupation*). Typically, production workers are considered to be unskilled and non-production workers are considered to be skilled. In other instances skilled and unskilled workers are frequently referred to as “white collars” and “blue collars.” Examples of the use of this proxy for skill can be found in Berndt & Christensen (1974), Kearney (1997), and Reshef (2007).

Data characteristics The studies in our sample also differ in the type of data used to produce estimates of the elasticity. An important aspect is data frequency. With higher frequencies, transitory variation from different frictions in the data 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 of labor demand 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 is 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 using linear aggregation techniques to aggregate nonlinear relationships. Even with the assumption of identical technologies in all firms, one cannot expect 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) warns 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 biases.

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. When a researcher wishes to generate estimates interpretable as short-run, she uses first-differences or an error-correction model, which require the time dimension to be reflected in the data. In case the long-run elasticity is estimated, 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 quite rare, and cross-sectional data generally enable greater disaggregation. To account for the potential negative trend in the elasticity over time as suggested earlier by Figure 1, we control for the midpoint of the data period employed in the primary study using the variable *Data midyear*. It should be noted that the literature typically discusses a possible recent increase in the elasticity, not a decrease (Autor, 2014). Finally, *Data length* captures the number of years of the data period used in the estimation of the elasticity.

Structural variation Inherent differences in the elasticity among countries, sectors, and genders 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).

Many studies use more homogenous samples to derive estimates that are not contaminated by structural variation (Ciccone & Peri, 2005). About 20% of the estimates in our sample are derived from male cohort studies. Borjas (2003) points out that female labor participation in the earlier cross-sections is much lower than male participation. Card & Lemieux (2001) analyze the substitutability of labor between age cohorts and argue that vast inter-cohort changes in female labor supply relevant to their sample period could contaminate the analysis. We account for these data differences with a variable called *Male workers*. 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-set (Berman *et al.*, 1994), may display 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 the multilevel (or nested) CES function, translog production function, and translog cost function. The one-level CES production function as in (2) was introduced by Solow (1956) but originally included capital and labor production factors (ignoring skill differences among workers). For the sake of simplicity, some authors consider solely (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 (*Multi-level 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.

Another type of the production function used by the primary studies in our sample is the transcendental logarithmic (translog) function. It is the second-order approximation of the CES function around $\rho = 0$ (Berndt & Christensen, 1973). Unlike in the CES framework, when the translog form is used the elasticity of substitution is allowed to vary with the relative supply of differently educated workers (Ciccone & Peri, 2005). The translog form has been used to define several measures of elasticity, including Allen-Uzawa, Morishima, and Shadow elasticity (Bergstrom & Panas, 1992). Most researchers working within the translog framework use a cost function instead of a production function. The main reason is convenience: with the translog cost function estimation of the elasticity is more straightforward.

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), changes in the level of education (Gallego, 2012; Autor *et al.*, 2008), 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, ethnicity, and capital stock.

Estimation techniques We control for all models that are *dynamic*, thus account for the fact that the elasticity may change in response to shocks (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 supply of workers or wages. As to the *short-run* and *long-run* elasticities, most researchers do not explicitly declare which type of elasticity they estimate. For the purpose of our analysis we code the estimated elasticity as short-run only if it is produced using first-difference estimation or the short-run equation of an error-correction model.

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 rarely available. 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. Silva (2008), on the other hand, instruments the relative share of skilled labor with a variable called “minimum wage intensity” defined as the share of workers potentially affected by new minimum wage legislation. To account for potential systematic variation stemming from the use of different estimation techniques, we codify estimates that are obtained using OLS, instrumental variable estimators, seemingly unrelated regression, and maximum likelihood.

Finally, we account for the fact that most of the collected elasticities are derived as inverse values of the estimated regression parameters, such as $-1/\sigma$ in (3). The “iron law of econometrics” implies that due to ubiquitous measurement error the estimated regression coefficients are biased towards zero (Hausman, 2001), thus biasing the implied elasticities upwards. Some studies switch the dependent and independent variable in order to estimate the elasticity directly, and the difference between direct and inverted estimates can be used to roughly gauge the importance of attenuation bias in the literature.

Publication characteristics To account for aspects of quality not captured by the variables introduced above, we employ three 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). Third, we distinguish between studies published in peer-reviewed outlets (*Published studies*) and other papers.

4.2 Estimation

Ideally we would regress the collected elasticities on the 37 variables described above using OLS. 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 each regression parameter. The method yields three basic statistics for each variable: posterior mean, posterior variance, and posterior inclusion probability. In our case BMA is to run 2^{37} regressions determined by all the possible combinations of the explanatory variables. We simplify this computationally demanding 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 particular 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 requires explicit priors concerning the model (model prior) and regression coefficients (g-prior). Our baseline model prior and g-prior reflects our lack of knowledge 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). Nevertheless, we perform several robustness checks

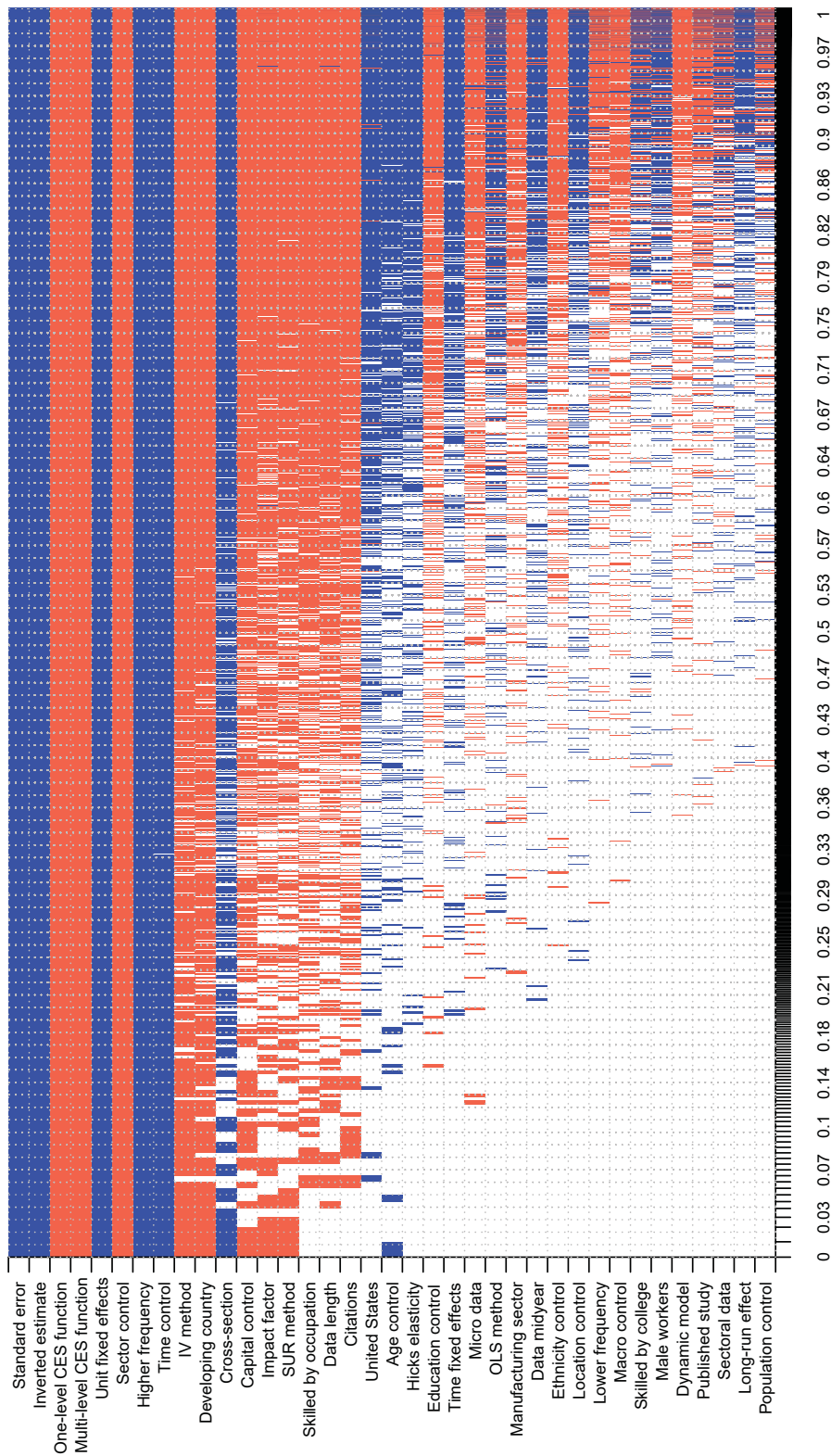
to make sure the results hold under different priors: we combine the unit information prior with the dilution prior according to George (2010), which accounts for collinearity in the data by adding a weight that is proportional to the determinant of the correlation matrix of the variables included in the individual model. Furthermore, 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 Mallows’s weights (Hansen, 2007) with orthogonalization of the covariate space according to Amini & Parmeter (2012) to narrow down the number of estimated models. For more details and applications of model averaging techniques in economics, we refer the reader to the superb survey by Steel (2020).

4.3 Results

The results of Bayesian model averaging are visualized in Figure 8. Each column represents an individual regression model, and the width of the column indicates the corresponding posterior model probability. The columns are ordered by posterior model probability from left to right in descending order. Each row of the figure represents a variable that we consider. The rows are ordered by the posterior inclusion probability from top to bottom in descending order. Each cell with a blue color (darker in grayscale) indicates a positive sign of the posterior mean of the regression coefficient for the variable in a given model. Each cell with a red color (lighter in grayscale) indicates a negative sign. If a variable is excluded from the model, the corresponding cell is blank. The figure suggests that approximately one third of our explanatory variables are useful in explaining the heterogeneity in the reported estimates of the elasticity of substitution; moreover, for these variables the signs of their coefficients are robust across virtually all the models.

The corresponding numerical results are reported in Table 6. The first specification represents our baseline BMA exercise. To interpret the posterior inclusion probabilities (PIPs) of the BMA means, we follow Jeffreys (1961), who denotes evidence of an effect that is ‘weak’ for

Figure 8: Model inclusion in Bayesian model averaging



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 6. All variables are described in Table 5.

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 6 represent robustness checks: first, ordinary least squares that exclude all the variables deemed unimportant by baseline 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, Appendix A provides more robustness checks that focus on different priors and weights (Table A3; including the dilution prior that addresses collinearity) and subsamples of our data (Table A4; long-run elasticities and Hicks elasticities). We will concentrate on the variables for which we have the most robust evidence across the models. The pre-eminent variable in this respect is the standard error, which shows the strongest association with the reported elasticity in all the models we run. Thus model averaging techniques corroborate our previous findings concerning publication bias.

Variable definition Although we include several distinct elasticity measures in our sample, including Hicks, Allen-Uzawa, Morishima, and Shadow, the different measures do not constitute an important factor driving the magnitude of the reported elasticities (in addition, we show in the third specification of Table A4 in Appendix A that our BMA results are robust to considering Hicks elasticities only). Several studies, including Goldin & Katz (2009), Mello (2011), and Binelli (2015) suggest that a lower education cut-off for skills (such as when skilled workers are defined as those with a high-school diploma) usually produces larger elasticities of substitution. The mechanism is intuitive but we do not find evidence that would support it. Nor do we support the notion that defining skills by occupation brings systematically different elasticities compared to definition by education, which was indicated by the *prima facie* pattern in Figure 5a.

Data characteristics Our results do suggest, however, that the differences between estimates can partly be explained by differences in data frequency. Higher than annual data granularity (usually monthly, quarterly, or semi-annual) systematically inflates the reported elasticity, which can be due to measurement error associated with higher frequency data. Measurement error in

Table 6: Why estimated elasticities vary

Response variable: Estimated elasticity of substitution	Bayesian model averaging			Frequentist check (OLS)			Frequentist model averaging		
	P. mean	P. SD	PIP	Coef.	SE	p-value	Coef.	SE	p-value
Constant	1.67	NA	1.00	1.22	0.41	0.00	1.86	1.41	0.19
Standard error (SE)	0.31	0.02	1.00	0.32	0.04	0.00	0.31	0.02	0.00
<i>Variable definition</i>									
Hicks elasticity	0.08	0.30	0.10				0.00	0.38	1.00
Skilled by college	0.00	0.04	0.03				0.00	0.31	1.00
Skilled by occupation	-0.17	0.25	0.39				-0.21	0.30	0.48
<i>Data characteristics</i>									
Higher frequency	0.98	0.25	1.00	1.04	0.40	0.01	0.90	0.36	0.01
Lower frequency	-0.01	0.07	0.03				0.00	0.44	1.00
Micro data	-0.03	0.12	0.08				0.00	0.43	1.00
Sectoral data	0.00	0.03	0.02				0.00	0.05	1.00
Cross-section	0.43	0.42	0.57	0.64	0.37	0.09	0.59	0.40	0.15
Data midyear	0.03	0.16	0.05				0.00	0.53	1.00
Data length	-0.07	0.11	0.37				-0.04	0.16	0.83
<i>Structural variation</i>									
United States	0.07	0.16	0.19				0.03	0.24	0.89
Developing country	-0.33	0.24	0.73	-0.36	0.17	0.03	-0.37	0.20	0.07
Male workers	0.00	0.04	0.03				0.00	0.11	1.00
Manufacturing sector	-0.02	0.10	0.06				0.00	0.07	1.00
<i>Design of the production function</i>									
One-level CES function	-1.75	0.30	1.00	-1.55	0.39	0.00	-1.65	0.50	0.00
Multi-level CES function	-1.20	0.31	1.00	-1.01	0.45	0.02	-1.09	0.61	0.07
Time control	0.63	0.18	0.99	0.69	0.19	0.00	0.52	0.23	0.02
Location control	0.01	0.06	0.04				0.00	0.11	1.00
Education control	-0.06	0.24	0.10				0.00	0.42	1.00
Macro control	-0.01	0.05	0.03				0.00	0.22	1.00
Population control	0.00	0.06	0.02				0.00	0.07	1.00
Sectoral control	-1.20	0.30	1.00	-1.19	0.33	0.00	-1.01	0.44	0.02
Age control	0.08	0.22	0.15				0.13	0.57	0.82
Ethnicity control	-0.02	0.14	0.05				0.00	0.24	1.00
Capital control	-0.32	0.39	0.47				-0.53	0.35	0.13
<i>Estimation technique</i>									
Dynamic model	0.00	0.03	0.02				0.00	0.09	1.00
Unit fixed effects	0.52	0.13	1.00	0.54	0.24	0.02	0.47	0.19	0.01
Time fixed effects	0.03	0.11	0.10				0.00	0.18	1.00
Long-run effect	0.00	0.04	0.02				0.00	0.07	1.00
OLS method	0.02	0.09	0.07				0.00	0.21	1.00
IV method	-0.38	0.25	0.78	-0.45	0.28	0.10	-0.43	0.34	0.21
SUR method	-0.46	0.64	0.39				-0.92	0.67	0.17
Inverted estimate	1.55	0.20	1.00	1.56	0.21	0.00	1.49	0.19	0.00
<i>Publication characteristics</i>									
Impact factor	-0.08	0.11	0.40				-0.16	0.11	0.15
Citations	-0.06	0.09	0.37				-0.03	0.05	0.63
Published study	0.00	0.03	0.02				0.00	0.26	1.00
Studies	61			61			61		
Observations	638			638			638		

Notes: P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. Bayesian model averaging employs the combination of the uniform model prior and the unit information g-prior recommended 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 level. Frequentist model averaging applies Mallows' 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 5. Additional details on the benchmark BMA exercise can be found in the Appendix A in Table A2 and Figure A1.

the independent variable results in attenuation bias in the estimated regression parameter, which translates to an upward bias in the elasticity (typically the inverse of the regression parameter). Next, our results suggest that data aggregation does not influence the reported elasticities systematically. Using purely cross-sectional data seem to affect the reported elasticity, but evidence for this effect is weak in both Bayesian and frequentist frameworks. It is also interesting to note that the downward trend in the reported elasticities shown in Figure 1 does not survive controlling for other aspects of study design. Moreover, the length of the sample period does not matter systematically for the resulting elasticity.

Structural variation Jones (2014) mentions that the differences in human capital between high- and low-skilled workers might be larger in rich countries because of higher standards in education and training. This notion constitutes an argument in favor of smaller elasticities of substitution in developed economies. But according to our results it is *developing countries* that show smaller elasticities. A significant portion of literature, including Klenow & Rodriguez-Clare (1997) and Caselli (2005), suggest that the efficiency of labor input varies between rich and poor countries. If we perceive the elasticity as a measure of ‘efficiency of a productive economy’ (de La Grandville, 1989), countries with a low efficiency of human capital display a smaller elasticity of substitution between skilled and unskilled labor. As Malmberg (2018) argues, larger elasticities imply that in order to maximize output efficiently the share of skilled labor within the force must be higher. The share of skilled labor is indeed much larger in developed countries than in the developing world. Moreover, we find that using only male labor force data, data for manufacturing, or data for the US does not systematically matter for the reported elasticities.

Design of the production function If the elasticity is estimated within a framework distinct from the CES (typically translog), it tends to be, with decisive posterior inclusion probability, larger. As Ciccone & Peri (2005) point out, for example, the main difference between the CES and translog approaches is that the translog one allows the elasticity of substitution to vary with the relative supply of differently skilled workers. If the wage share of more skilled workers increases over time, in the translog specification the elasticity is pushed upward. While translog is more flexible and allows for the estimation of elasticities between different factors, it has the

incremental disadvantage of being more prone to misspecification bias and measurement error (see, for example, Freeman & Medoff, 1982).

Next, we observe that controlling for the time period is associated with systematically larger elasticities. Functions of time are typically added to regression specifications in order to proxy for varying demand for skills, and a lack of this control may constitute omitted variable bias. On the other hand, once the authors of primary studies control for sectoral differences in the data, the elasticity estimates get systematically smaller. We also find weak evidence across our models that the inclusion of capital control matters: in theory, industries with a larger stock of capital have higher wages because labor productivity is higher. Increases in the capital stock increase the marginal product of skilled labor, but decrease the marginal product of unskilled labor; excluding the capital control could thus bias the elasticity estimates.

Estimation technique When studies employ panel data, fixed effects can be used to control for unobserved features that might affect the education of workers or the technology chosen by firms. Our results suggest that ignoring unit fixed-effects can drive the reported elasticity downwards. Time fixed-effects, on the other hand, do not affect the estimates systematically. Next, we do not find that long- and short-run effects differ systematically. In contrast, we find that accounting for endogeneity yields systematically smaller elasticity estimates. The evidence for this effect gets even stronger once only Hicks or long-run elasticity estimates are considered (see Table A4). Our results are also consistent with attenuation bias, which drives the regression estimate downwards and the elasticity, computed typically as the estimate’s inverse, upwards. The result is in line with Bowles (1970), who nevertheless argues that the bias is not large. In our case, however, direct estimates tend to be 1.5 smaller than inverted estimates, which implies a large effect.

Publication characteristics Our results indicate that publication characteristics, which we included as rough proxies for quality not captured by the remaining variables, are not systematically associated with the reported elasticities. Influential and obscure studies tend to present similar results on average once the characteristics of data and methods are taken into account.

The discussion has so far focused mostly on statistical significance, especially the posterior inclusion probability resulting from Bayesian model averaging. Table 7 explores the economic

significance of the key variables identified as important by BMA and FMA. The table shows the effect on the implied elasticity when we increase the value of the corresponding variable by one standard deviation (the panel on the left) and when we increase the value of the corresponding variable from the sample minimum to sample maximum (the panel on the right). The variables with the largest effect on the substitution elasticity are the standard error (a proxy for publication bias), functional form (CES versus translog), and inverted estimation (a proxy for attenuation bias). Changes in these characteristics can routinely alter the reported elasticity by 0.7 or more.

Table 7: Economic significance of key variables

	One-std.-dev. change		Maximum change	
	Effect on σ	% of best practice	Effect on σ	% of best practice
Standard error	1.08	151%	5.06	705%
Higher frequency	0.25	35%	0.98	137%
Cross-section	0.17	24%	0.43	60%
Developing country	-0.15	-22%	-0.33	-47%
One-level CES function	-0.87	-121%	-1.75	-244%
Multi-level CES function	-0.53	-74%	-1.20	-168%
Time control	0.32	44%	0.63	88%
Sectoral control	-0.25	-34%	-1.20	-167%
Unit fixed effects	0.26	37%	0.52	73%
IV method	-0.16	-23%	-0.38	-53%
Inverted estimate	0.77	108%	1.55	216%

Notes: The table shows *ceteris paribus* changes in the reported elasticities implied by changes in the variables that reflect the context in which researchers obtain their estimates. For example, increasing the estimate's standard error by one standard deviation is associated with an increase in the estimated elasticity by 1.08, about one and half times the size of the implied best practice estimate (conditional on ideal data, method, and publication characteristics, as described in Table 8). Increasing the standard error from the sample minimum to the sample maximum is associated with an increase in the estimated elasticity by 5.06, about seven times the best practice estimate. A detailed description of the variables is available in Table 5.

The model averaging analysis suggests that the reported elasticity of substitution is affected by many data and method choices, including ones that arguably lead to a bias. Publication and attenuation bias work in concert, but the other potential biases often influence the elasticity in the opposite direction. As the bottom line of our analysis we thus compute an implied elasticity conditional on the entire sample of collected elasticities, our baseline BMA results, and a definition of best practice methodology in the literature. Since best practice is subjective, we choose three alternative strategies. First and second, we rely on two distinct definitions from the literature: Card (2009) and Autor (2014). These are meticulous contributions by leaders in the field and have also been published in prestigious journals, *American Economic Review*

and *Science*. We copy their data and method characteristics and plug them in the values of our variables in order to compute the fitted values from BMA and, hence, the implied elasticity. The only exception are proxies for publication and attenuation bias in the literature: we plug zero for the standard error and 0.5 for the dummy variable that equals one for inverted estimates. Third, we create a subjective definition of best practice based on our reading of the literature. It turns out that the results of the three strategies do not differ much.

Table 8: Best-practice elasticities implied for different data sets

	Subjective best practice	Autor (2014)	Card (2009)
All data	0.72 (0.21, 1.22)	0.53 (0.21, 0.85)	0.49 (-0.51, 1.49)
Manufacturing & male workers	0.53 (0.01, 0.91)	0.48 (0.18, 0.79)	0.44 (-0.56, 1.44)
USA	0.86 (0.3, 1.42)	0.68 (0.37, 0.98)	0.64 (-0.37, 1.64)
Developing countries	0.46 (0.01, 0.91)	0.27 (-0.16, 0.71)	0.23 (-0.81, 1.28)

Notes: The table presents mean estimates of the elasticity of substitution between skilled and unskilled labor implied by the Bayesian model averaging and our definition of best practice in the first column, the methods used by Autor (2014) in the second column and by Card (2009) in the third column (with the exception that in the latter two cases we plug in 0.5 instead of 1 for the variable *Inverted estimate* and zero for the variable *Standard error*). That is, the table attempts to answer the question what the mean elasticities would look like if the literature was approximately corrected for publication and attenuation biases and all studies in the literature used the same strategy as the one that we prefer or the ones employed by Autor (2014) and Card (2009). The 95% confidence intervals in parentheses are approximate and constructed using the standard errors estimated by OLS with standard errors clustered at the study level.

Our subjective definition of best practice is the following. We plug in zero for the standard error in order to correct for publication bias. (Note that in Section 3 we have shown that the linear correction for publication bias is conservative.) We prefer long-run estimates of the elasticity, because the long-run estimate is more relevant for calibration. We also prefer firm-level panel data, annual granularity, and maximum length of the data period. We plug in the maximum for the midyear of the data used in the primary studies because we want to give more weight to recent data. We prefer when skills are defined by a college degree. 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. We also prefer studies published in peer-reviewed journals with a large 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 8 reports the results. The first row shows the overall estimate, the second row shows the estimate corresponding to a more homogeneous dataset of male workers in manufacturing, the third row shows the estimate for the US, and the last row shows the estimate for developing countries. We find it useful to focus on the estimate for the US because it is comparable with the best known estimates in the literature: our analysis implies an elasticity of 0.6–0.9, much less than the typical calibration of 1.5. The confidence intervals are wide due to the large number of variables included in model averaging, but a similar point estimate of the implied elasticity with tight confidence intervals can be derived for the US when one runs a parsimonious regression controlling only for publication bias, attenuation bias, and (downward) omitted variable bias stemming from ignoring the demand for skills. In a nutshell, we find that the literature is consistent with the notion that skilled and unskilled labor are gross complements.

5 Concluding Remarks

We collect 729 estimates of the elasticity of substitution between skilled and unskilled labor reported in 76 studies to provide the first quantitative synthesis of the literature. We measure the extent of two biases that drive the mean reported elasticity upwards: publication bias (stemming from the underreporting of small estimates) and attenuation bias (stemming from measurement error and estimating the inverse of the elasticity in most regressions). Correcting for publication bias by itself shrinks the mean elasticity from 1.9 to 1, suggesting that selective reporting is associated with a twofold exaggeration. The twofold exaggeration emerges as a rule of thumb in economics and social sciences more generally, as evidenced by both large-scale surveys of meta-analyses (Doucouliagos & Stanley, 2013; Ioannidis *et al.*, 2017) and extensive replication studies (Klein *et al.*, 2014; Open Science Collaboration, 2015; Camerer *et al.*, 2018; Klein *et al.*, 2018). Direct estimates of the elasticity are on average 1.5 smaller than elasticities for which the regression coefficient needs to be inverted, a finding that is consistent with attenuation bias in the vicinity of 0.75. 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). Unfortunately in the case of the elasticity of substitution the two biases work in concert.

The aforementioned results hold when we control for additional 35 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 (which is the natural response to model uncertainty in the Bayesian setting; Steel, 2020) and frequentist model averaging. For the former we also employ a dilution prior that addresses potential collinearity (George, 2010). We find that larger estimated elasticities are associated with data from developed countries, higher frequency, translog function, ignoring endogeneity, and controlling for the demand for skills. We then compute the implied elasticity conditional on best practice methodology (using Card, 2009, and Autor, 2014, as benchmarks). The resulting estimates for the US range between 0.6 and 0.9, suggesting that skilled and unskilled labor form gross complements. Among other things, this finding implies that in the canonical model skill-biased technical change compresses the skill premium (Johnson, 1997).

Three qualifications of our results are in order. First, our correction for attenuation bias is crude: we compare the estimates of the elasticity when the dependent and independent variables are switched in the regression. Such approach gives us bounds for the interval of plausible estimates, but does not clearly identify the underlying elasticity. We use the midpoint of the interval. To measure the extent of attenuation bias precisely, one would need to collect the data used in the 76 primary studies. Such an extension, while relevant, is on the boundary of feasibility and beyond our scope in this paper. Second, our correction for publication bias (following Stanley, 2005; Ioannidis *et al.*, 2017; Andrews & Kasy, 2019; Furukawa, 2020) relies on the assumption that in the absence of the bias no correlation exists between estimates and their standard errors. While common in the meta-analysis literature, the assumption can be violated in labor economics: it is enough for an unobserved method choice to influence both estimates and standard errors in the same direction, and a positive correlation arises that will be spuriously interpreted as publication bias. A straightforward solution is to use the number of observations as an instrument for the standard error, but unfortunately in our case the instrument proves weak. We thus additionally employ the caliper test (Gerber & Malhotra, 2008; Brodeur *et al.*, 2020), which does not rely on the assumption and also finds that negative

and insignificant estimates are underreported. Third, estimates reported in the same study are not independent. We address this issue by clustering standard errors at the study level, using wild bootstrap, and employing a Hierarchical Bayesian model.

References

- ACEMOGLU, D. (2002): “Technical change, inequality, and the labor market.” *Journal of Economic Literature* **40**(1): pp. 7–72.
- ALLEN, G. D. & J. R. HICKS (1934): “A Reconsideration of the Theory of Value, Part II.” *Economica* **1**: pp. 196–219.
- AMINI, S. M. & C. F. PARMETER (2012): “Comparison of model averaging techniques: Assessing growth determinants.” *Journal of Applied Econometrics* **27**(5): pp. 870–876.
- ANDREWS, I. & M. KASY (2019): “Identification of and correction for publication bias.” *American Economic Review* **109**(8): pp. 2766–2794.
- ANGRIST, J. D. (1995): “The economic returns to schooling in the West Bank and Gaza Strip.” *The American Economic Review* **85**(5): pp. 1065–1087.
- 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.
- 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 (1973): “The translog function and the substitution of equipment, structures, and labor in US manufacturing 1929–68.” *Journal of Econometrics* **1**(1): pp. 81–113.
- 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. MORRISON (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.” *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.
- 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. & 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.
- 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* (forthcoming).
- 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.
- BRUNS, S. B. & J. P. A. IOANNIDIS (2016): “p-Curve and p-Hacking in Observational Research.” *PLoS ONE* **11**(2): p. e0149144.
- CAMERER, C., A. DREBER, F. HOLZMEISTER, T. H. HO, J. HUBER, M. JOHANNESSON, M. KIRCHLER, G. NAVE, B. A. NOSEK, & T. PFEIFFER (2018): “Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015.” *Nature Human Behaviour* **2**: pp. 637–644.
- CANTORE, C., F. FERRONI, & M. A. LEÓN-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. & A. B. KRUEGER (1995): “Time-series minimum-wage studies: A meta-analysis.” *American Economic Review* **85**(2): pp. 238–243.
- 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.
- CASELLI, F. (2005): “Accounting for Cross-Country Income Differences.” In P. AGHION & S. DURLAUF (editors), “Handbook of Economic Growth,” volume 1, Part A, chapter 9, pp. 679–741. Elsevier, 1 edition.
- 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.
- CRUZ, M., E. MILET, & M. OLARREAGA (2017): “Online exports and the wage gap.” *Policy Research Working Paper Series 8160/2017*, Trade and Competitiveness Global Practice Group, The World Bank.
- 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., 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.
- 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.
- EGGER, M., G. D. SMITH, M. SCHNEIDER, & C. MINDER (1997): “Bias in meta-analysis detected by a simple, graphical test.” *BMJ* **315**(7109): pp. 629–634.
- EICHER, T. S., C. PAPAGEORGIOU, & A. E. RAFTERY (2011): “Default priors and predictive performance in Bayesian model averaging, with application to growth determinants.” *Journal of Applied Econometrics* **26**(1): pp. 30–55.
- 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 im-

- pect 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 (2017): “Skill premium, labor supply and changes in the structure of wages in Latin America.” *IDB Working Paper Series 786/2017*, Inter-American Development Bank.
- 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. (1986): “Demand for education.” In O. ASHENFELTER & R. LAYARD (editors), “Handbook of Labor Economics,” volume 1, chapter 6, pp. 357–386. Elsevier Science Publishers BV.
- FREEMAN, R. B. & J. L. MEDOFF (1982): “Substitution between production labor and other inputs in unionized 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.
- 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. (2015): “Can job displacements explain the UK’s productivity puzzle?” *Technical report*, University of Southampton.
- GLITZ, A. & D. WISSMANN (2017): “Skill premiums and the supply of young workers in Germany.” *IZA Discussion Papers 10901*, Institute of Labor Economics.
- GOLDIN, C. D. & L. F. KATZ (2009): *The race between education and technology*. Cambridge, MA: Harvard University Press, 1 edition.
- GRILICHES, Z. (1977): “Estimating the returns to schooling: Some econometric problems.” *Econometrica* **45(1)**: pp. 1–22.
- GYIMAH-BREMPPONG, 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. (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.
- 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.
- HEDGES, L. V. (1992): “Modeling Publication Selection Effects in Meta-Analysis.” *Statistical Science* **7(2)**: pp. 246–255.
- HICKS, J. (1963): *The theory of wages*. London: Macmillan, 2 edition.
- IMAI, T., T. A. RUTTER, & C. F. CAMERER (2020): “Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets.” *Economic Journal* (**forthcoming**).
- IOANNIDIS, J. P., T. D. STANLEY, & H. DOUCOULIAGOS (2017): “The Power of Bias in Economics Research.” *Economic Journal* **127(605)**: pp. F236–F265.
- JAMET, S. (2005): “Allegements generaux de cotisations sociales et emploi peu qualifie: De l’impact sectoriel a l’effet macro-economique.” *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.
- JOHNSON, G. E. (1970): “The demand for labor by

- educational category.” *Southern Economic Journal* **37(2)**: pp. 190–204.
- JOHNSON, G. E. (1997): “Changes in Earnings Inequality: The Role of Demand Shifts.” *Journal of Economic Perspectives* **11(2)**: pp. 41–54.
- JONES, B. F. (2014): “The Human Capital Stock: A Generalized Approach.” *American Economic Review* **104(11)**: pp. 3752–3777.
- 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**: pp. 29–50.
- KEARNEY, I. (1997): “Estimating the demand for skilled labour, unskilled labour and clerical workers: A dynamic framework.” *ESRI Working Paper Series 91/1997*, 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.
- KLEIN, R. A., K. A. RATLIFF, M. VIANELLO, R. B. ADAMS, S. BAHNIK, M. J. BERNSTEIN, . . . , & B. A. NOSEK (2014): “Investigating variation in replicability: A ‘many labs’ replication project.” *Social Psychology* **45(3)**: pp. 142–152.
- KLEIN, R. A., M. VIANELLO, F. HASSELMAN, B. G. ADAMS, R. B. ADAMS, S. ALPER, . . . , Z. ZHIJA, & B. A. NOSEK (2018): “Many Labs 2: Investigating Variation in Replicability Across Samples and Settings.” *Advances in Methods and Practices in Psychological Science* **1(4)**: pp. 443–490.
- KLENOW, P. J. & A. RODRIGUEZ-CLARE (1997): “The neoclassical revival in growth economics: Has it gone too far?” *NBER Macroeconomics Annual* **17**: pp. 73–114.
- KLOTZ, B., R. MADDOO, & 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.
- 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.
- DE LA GRANDVILLE, O. (1989): “In quest of the slutzky diamond.” *American Economic Review* **79(3)**: pp. 468–81.
- LEY, E. & M. F. STEEL (2009): “On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression.” *Applied Econometrics* **24**: pp. 651–674.
- 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., C. SÁNCHEZ-PÁRAMO, & 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.
- 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 Economía” 614*, Banco de la Republica de Colombia.
- MELLO, M. (2011): “Another look at panel estimates of the elasticity of substitution.” *Working paper*, IB-MEC 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.
- MORISHIMA, M. (1967): “A few suggestions on the theory of elasticity (in Japanese).” *Keizai Hyoron (Economic Review)* **16**: pp. 144–150.
- 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.
- 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.
- OPEN SCIENCE COLLABORATION (2015): “Estimating the reproducibility of psychological science.” *Science*

- 349(6251)**. Article 4716.
- 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. HOETING (1997): "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* **92**: 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.
- 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." *OECD Development Centre Working Papers 119*, OECD Development Centre, 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." *Queen's Economics Department Working Paper 1406*, Department of Economics, Queen's University, Canada: Kingston.
- ROSSI, P. E., G. M. ALLENBY, & R. MCCULLOCH (2005): "Markov Chain Monte Carlo Methods." In V. BARNETT, J. S. HUNTER, & D. G. KENDALL (editors), "Bayesian Statistics and Marketing," volume 1 of *Wiley Series in Probability and Statistics*, chapter 3, pp. 49–102. John Wiley & Sons, Ltd.
- SANTAMARIA, M. (2004): "Income inequality, skills and trade: Evidence from Colombia during the 80s and 90s." *Documento CEDE 2004-02*, Bogota: Universidad de los Andes.
- SILVA, J. C. (2008): "Young employment, job-skill composition and minimum wages: Evidence from a "natural experiment"." *NIPE Working paper 4/2008*, Departamento de Economia, Universidade do Minho, Portugal.
- SOLOW, R. M. (1956): "A contribution to the Theory of economic growth." *The Quarterly Journal of Economics* **70(1)**: pp. 65–94.
- STANLEY, T. D. (2001): "Wheat from Chaff: Meta-analysis as Quantitative Literature Review." *Journal of Economic Perspectives* **15(3)**: pp. 131–150.
- STANLEY, T. D. (2005): "Beyond Publication Bias." *Journal of Economic Surveys* **19(3)**: pp. 309–345.
- STANLEY, T. D., S. B. JARRELL, & H. DOUCOULIAGOS (2010): "Could It Be Better to Discard 90% of the Data? A Statistical Paradox." *The American Statistician* **64(1)**: pp. 70–77.
- STEEL, M. F. J. (2020): "Model Averaging and its Use in Economics." *Journal of Economic Literature* (**forthcoming**).
- TINBERGEN, J. (1974): "Substitution of graduate by other labour." *Kyklos* **27(2)**: pp. 217–226.
- UZAWA, H. (1962): "Production functions with Constant Elasticities of Substitution." *Review of Economic Studies* **29**: pp. 291–299.
- 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**: pp. 131–144.
- WEI, X., Z. GUAN, G. ONEL, & F. ROKA (2016): "Imperfect substitution between immigrant and native farm workers in the United States." *Paper prepared for presentation at the 2016 Agricultural & Applied Economics Association Annual Meeting (during July 31-August 2 in Boston, Massachusetts) 235958*, Agricultural and Applied Economics Association.
- WELCH, F. (1970): "Education in production." *Journal of Political Economy* **78(1)**: pp. 35–59.
- YANG, D. (2012): "The responsiveness of firms to labor market changes in China: Evidence from micro level data." *Paper prepared for presentation at the 2012 4th CIER/IZA Annual Workshop on Research in Labor Economics, (during July 19-21 at IZA, Bonn) 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.

Appendices

A Robustness Checks and BMA Diagnostics

Table A1: Linear and nonlinear techniques detect publication bias among long-run estimates

<i>Panel A: unweighted</i>	OLS	Fixed effects	Hierarchical Bayes
Standard error (<i>publication bias</i>)	0.375 ^{***} (0.0464) [0.209, 0.459]	0.312 ^{***} (0.0393) –	0.470 ^{***} (0.0045) –
Constant (<i>effect beyond bias</i>)	1.306 ^{***} (0.300) [0.646, 2.063]	1.403 ^{***} (0.0604) –	1.440 ^{***} (0.0047) –
Observations	600	600	600
<i>Panel B: weighted</i>	Study-weighted	Precision-weighted	Top journals
Standard error (<i>publication bias</i>)	0.256 ^{***} (0.0901) [0.048, 0.503]	0.782 ^{***} (0.143) [0.486, 1.140]	0.376 ^{***} (0.109) [-0.115, 1.109]
Constant (<i>effect beyond bias</i>)	1.791 ^{***} (0.233) [1.338, 2.267]	1.002 ^{**} (0.483) [-0.127, 2.165]	1.564 ^{***} (0.283) [0.845, 3.263]
Observations	600	600	149
<i>Panel C: non-linear</i>	WAAP (Ioannidis <i>et al.</i> , 2017)	Stem-based method (Furukawa, 2020)	Selection model (Andrews & Kasy, 2019)
Effect beyond bias	0.954 ^{***} (0.054)	0.829 ^{***} (0.045)	0.963 ^{***} (0.059)
Observations	600	600	600

Notes: Only long-run estimates are considered. The first two panels report the results of the regression $\sigma_{ij} = \sigma_0 + \beta \cdot SE(\sigma_{ij}) + \epsilon_{ij}$, where σ_{ij} denotes the i -th effect estimated in the j -th study, and $SE(\sigma_{ij})$ denotes its standard error. In the second column of Panel A study-level fixed effects are included. In the first column of Panel B the model is weighted by the inverse of the number of estimates per study. Precision-weighting indicates that the model is weighted by the inverse of the standard error of an estimate. The “top journals” subsample includes the most prestigious outlets in our sample: American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Econometrica, Review of Economics and Statistics, and Science. In Panel C, WAAP stands for Weighted Average of the Adequately Powered. Standard errors, clustered at the study level, are reported in parentheses (except for Hierarchical Bayes, which has posterior standard deviation in parentheses; stars for Hierarchical Bayes are presented only as an indication of the parameter’s importance to keep visual consistency with the rest of the table). 95% confidence intervals from wild bootstrap in square brackets (Roodman *et al.*, 2018). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Summary of the benchmark BMA estimation

<i>Mean no. regressors</i>	<i>Draws</i>	<i>Burn-ins</i>	<i>Time</i>	<i>No. models visited</i>
14.1637	$3 \cdot 10^5$	$1 \cdot 10^5$	1.229579 mins	74,645
<i>Modelspace</i>	<i>Visited</i>	<i>Topmodels</i>	<i>Corr PMP</i>	<i>No. obs.</i>
$1.4 \cdot 10^{11}$	22.10%	100%	0.9561	638
<i>Model prior</i>	<i>g-prior</i>	<i>Shrinkage-stats</i>		
Uniform/18.5	UIP	$A_v = 0.9984$		

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 A1: Benchmark model size and convergence, benchmark BMA

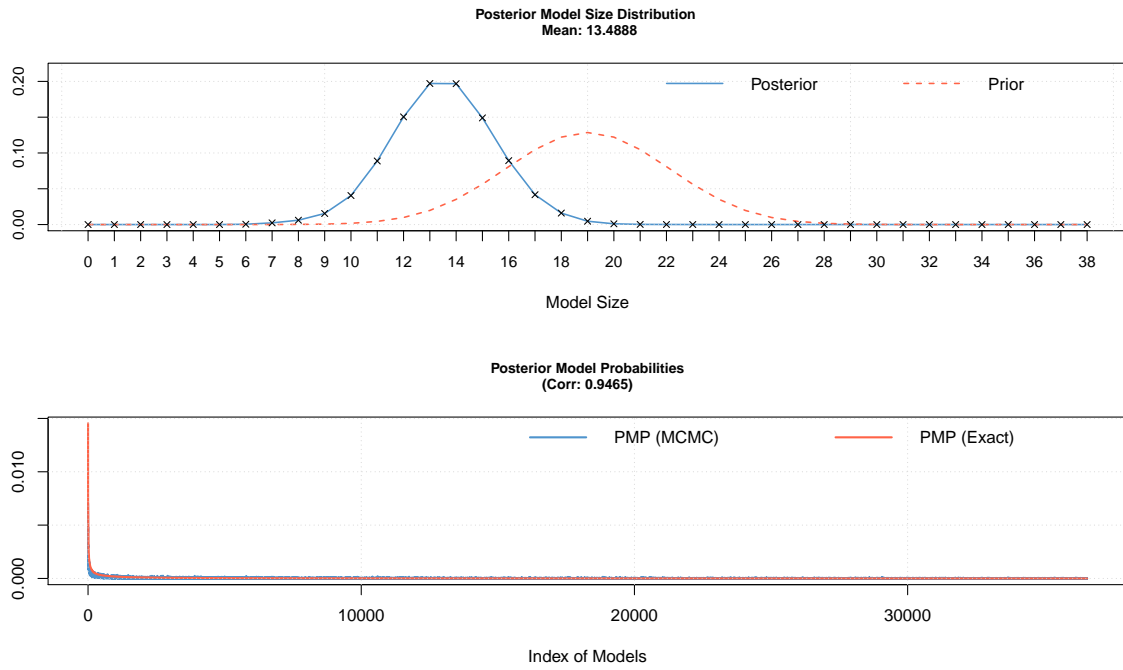


Table A3: Why elasticities vary (alternative priors and weights)

Response variable: Estimated elasticity	Bayesian model averaging (dilution prior)			Bayesian model averaging (hyper-g prior)			Bayesian model averaging (weighted)		
	P. mean	P. SD	PIP	P. mean	P. SD	PIP	P. mean	P. SD	PIP
Constant	1.54	NA	1.00	1.88	NA	1.00	0.09	NA	1.00
Standard error (SE)	0.31	0.02	1.00	0.31	0.02	1.00	0.02	0.00	0.99
<i>Variable definition</i>									
Hicks elasticity	0.07	0.27	0.08	0.08	0.29	0.12	0.76	0.13	1.00
Skilled by college	0.00	0.03	0.01	0.00	0.06	0.06	0.13	0.07	0.83
Skilled by occupation	-0.16	0.25	0.34	-0.18	0.24	0.45	0.00	0.02	0.05
<i>Data characteristics</i>									
Higher frequency	0.88	0.30	0.96	0.98	0.26	1.00	0.02	0.05	0.15
Lower frequency	-0.01	0.08	0.03	-0.01	0.08	0.07	-0.02	0.05	0.16
Micro data	-0.01	0.08	0.03	-0.05	0.16	0.15	0.00	0.02	0.06
Sectoral data	0.00	0.02	0.01	0.00	0.04	0.04	-0.42	0.04	1.00
Cross-section	0.37	0.39	0.53	0.47	0.45	0.61	0.00	0.03	0.05
Data midyear	0.02	0.11	0.03	0.03	0.18	0.08	0.23	0.04	1.00
Data length	-0.05	0.09	0.28	-0.09	0.12	0.44	0.00	0.00	0.03
<i>Structural variation</i>									
United States	0.04	0.13	0.13	0.05	0.15	0.17	0.00	0.01	0.03
Developing country	-0.22	0.24	0.52	-0.41	0.22	0.86	0.00	0.01	0.03
Male workers	0.00	0.03	0.01	0.00	0.04	0.05	0.00	0.02	0.06
Manufacturing sector	-0.01	0.08	0.04	-0.02	0.10	0.08	0.00	0.02	0.05
<i>Design of the production function</i>									
One-level CES function	-1.69	0.29	1.00	-1.81	0.31	1.00	-1.33	0.09	1.00
Multi-level CES function	-1.19	0.31	0.99	-1.24	0.31	1.00	-1.16	0.09	1.00
Time control	0.63	0.19	0.97	0.61	0.17	1.00	0.34	0.05	1.00
Location control	0.00	0.04	0.02	0.02	0.08	0.08	0.00	0.01	0.03
Education control	-0.03	0.17	0.05	-0.12	0.31	0.18	0.37	0.13	0.96
Macro control	-0.01	0.06	0.03	-0.01	0.06	0.05	0.37	0.08	1.00
Population control	0.00	0.05	0.01	0.00	0.09	0.03	0.46	0.23	0.88
Sectoral control	-1.12	0.35	0.97	-1.22	0.30	1.00	0.00	0.01	0.03
Age control	0.03	0.14	0.06	0.19	0.31	0.34	0.01	0.04	0.07
Ethnicity control	-0.01	0.10	0.03	-0.04	0.16	0.09	0.01	0.05	0.05
Capital control	-0.16	0.32	0.24	-0.49	0.38	0.73	0.00	0.01	0.03
<i>Estimation technique</i>									
Dynamic model	0.00	0.02	0.01	-0.01	0.04	0.05	0.01	0.02	0.10
Unit fixed effects	0.49	0.15	0.97	0.53	0.13	1.00	0.00	0.01	0.05
Time fixed effects	0.01	0.08	0.04	0.05	0.13	0.16	-0.02	0.05	0.15
Long-run effect	0.00	0.03	0.01	0.00	0.06	0.04	-0.05	0.09	0.32
OLS method	0.02	0.08	0.06	0.00	0.08	0.07	-0.31	0.06	1.00
IV method	-0.25	0.26	0.54	-0.48	0.21	0.93	-0.38	0.06	1.00
SUR method	-0.20	0.48	0.18	-0.81	0.69	0.67	-0.04	0.12	0.14
Inverted estimate	1.53	0.20	1.00	1.54	0.20	1.00	0.03	0.06	0.22
<i>Publication characteristics</i>									
Impact factor	-0.04	0.08	0.20	-0.14	0.12	0.64	0.05	0.03	0.86
Citations	-0.04	0.08	0.24	-0.06	0.09	0.37	0.00	0.01	0.04
Published study	0.00	0.02	0.01	0.00	0.04	0.04	-0.21	0.05	1.00
Studies	61			61			61		
Observations	638			638			638		

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 (BMA) using uniform model prior (Eicher *et al.*, 2011) and dilution prior suggested by George (2010), which accounts for collinearity. The second specification uses a random model prior advocated by Ley & Steel (2009) and the data-dependent hyper-g prior suggested by Feldkircher & Zeugner (2012). In the third specification we weight our data by the inverse of the number of observations used in a study. All variables are described in Table 5.

Table A4: Why elasticities vary (different subsamples)

Response variable: Estimated elasticity	Bayesian model averaging (long-run elasticity)			Frequentist model averaging) (long-run elasticity)			Bayesian model averaging (Hicks elasticity)		
	P. mean	P. SD	PIP	Mean	SE	p-val.	P. mean	P. SD	PIP
Constant	1.61	NA	1.00	2.12	2.20	0.34	1.60	NA	1.00
Standard error (SE)	0.31	0.02	1.00	0.31	0.02	0.00	0.31	0.02	1.00
<i>Variable definition</i>									
Hicks elasticity	0.03	0.19	0.06	0.00	0.20	1.00			
Skilled by college	0.03	0.11	0.11	0.00	0.29	1.00	0.00	0.03	0.03
Skilled by occupation	-0.07	0.18	0.18	-0.15	0.31	0.62	-0.07	0.16	0.18
<i>Data characteristics</i>									
Higher frequency	0.87	0.42	0.88	0.86	0.44	0.05	1.17	0.25	1.00
Lower frequency	-0.06	0.17	0.13	0.00	0.61	1.00	-0.01	0.07	0.04
Micro data	-0.02	0.11	0.06	0.00	0.43	1.00	-0.02	0.10	0.07
Sectoral data	0.00	0.03	0.02	0.00	0.10	1.00	0.00	0.03	0.03
Cross-section	0.10	0.26	0.18	0.41	0.40	0.31	0.65	0.45	0.75
Data midyear	0.16	0.39	0.19	0.00	0.71	1.00	0.09	0.29	0.12
Data length	-0.03	0.08	0.21	-0.06	0.13	0.68	-0.06	0.11	0.26
<i>Structural variation</i>									
United States	0.01	0.07	0.05	0.03	0.17	0.88	0.08	0.17	0.21
Developing country	-0.52	0.18	0.96	-0.41	0.23	0.07	-0.36	0.24	0.77
Male workers	0.01	0.06	0.04	0.00	0.07	1.00	0.00	0.03	0.03
Manufacturing sector	-0.01	0.06	0.04	0.00	0.02	1.00	-0.02	0.10	0.06
<i>Design of the production function</i>									
One-level CES function	-1.82	0.30	1.00	-1.72	0.43	0.00	-2.00	0.33	1.00
Multi-level CES function	-1.18	0.33	1.00	-1.16	0.54	0.03	-1.41	0.34	1.00
Time control	0.26	0.26	0.60	0.42	0.20	0.04	0.63	0.18	0.99
Location control	0.00	0.06	0.03	0.00	0.05	1.00	0.01	0.07	0.05
Education control	-0.01	0.12	0.04	0.00	0.33	1.00	-0.16	0.37	0.19
Macro control	-0.01	0.05	0.04	0.00	0.23	1.00	0.00	0.04	0.03
Population control	-0.01	0.10	0.03	0.00	0.28	1.00	0.00	0.07	0.02
Sectoral control	-1.18	0.31	1.00	-1.02	0.44	0.02	-0.78	0.49	0.80
Age control	0.22	0.37	0.31	0.18	0.66	0.78	0.04	0.15	0.08
Ethnicity control	0.00	0.07	0.03	0.00	0.11	1.00	-0.03	0.16	0.06
Capital control	-0.89	0.43	0.89	-0.66	0.41	0.11	-0.22	0.34	0.35
<i>Estimation technique</i>									
Dynamic model	0.00	0.03	0.02	0.00	0.09	1.00	-0.01	0.06	0.05
Unit fixed effects	0.44	0.15	0.97	0.41	0.19	0.03	0.53	0.13	1.00
Time fixed effects	0.02	0.11	0.07	0.00	0.24	1.00	0.03	0.10	0.09
Long-run effect							0.00	0.05	0.03
OLS method	0.03	0.13	0.08	0.00	0.22	1.00	0.00	0.06	0.04
IV method	-0.59	0.23	0.94	-0.52	0.35	0.13	-0.50	0.21	0.93
SUR method	-0.64	0.70	0.52	-1.01	0.70	0.15	0.00	0.11	0.02
Inverted estimate	1.61	0.20	1.00	1.57	0.19	0.00	1.56	0.19	1.00
<i>Publication characteristics</i>									
Impact factor	-0.05	0.10	0.21	-0.07	0.09	0.44	-0.12	0.12	0.56
Citations	-0.16	0.11	0.76	-0.16	0.11	0.14	-0.02	0.06	0.16
Published study	0.00	0.04	0.03	0.00	0.31	1.00	0.00	0.03	0.03
Studies	58			58			57		
Observations	600			600			620		

Notes: P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. In the first specification from the left we employ Bayesian model averaging (BMA) on the subsample of long-run elasticities using a uniform model prior with the unit information prior on Zellner's g (Eicher *et al.*, 2011). The specification yields results almost identical to that using the dilution prior suggested by George (2010), which accounts for collinearity. The second specification employs frequentist model averaging by applying Mallows' weights (Hansen, 2007) using orthogonalization of covariate space suggested by Amini & Parmeter (2012) to reduce the number of estimated models. In the third specification we only use Hicks elasticities and apply uniform model prior with the unit information prior on Zellner's g (Eicher *et al.*, 2011). All variables are described in Table 5.