CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Ability bias in the returns to schooling: How large it is and why it matters

Diploma Thesis

Author: Petr Čala

Study program: Economics and Finance

Supervisor: doc. PhDr. Zuzana Havránková, Ph.D.

Year of defense: 2024

Declaration of Authorship

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

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

Prague, December 10, 2023

Petr Cala

Abstract

Abstract text heere

JEL Classification	TBA
Keywords	TBA

Title	Ability bias in the returns to schooling: How large it is and why it matters
Author's e-mail	61505008@fsv.cuni.cz
Supervisor's e-mail	zuzana.havrankova@fsv.cuni.cz

Abstrakt

Abstract text here, in Czech

Klasifikace JEL	TBA
Klíčová slova	TBA
Název práce	Dovednostní zkreslení v návratnosti do vzdělání: Jak velké je a proč na tom záleží?
E-mail autora	61505008@fsv.cuni.cz
E-mail vedoucího práce	zuzana.havrankova@fsv.cuni.cz

Acknowledgments

Typeset in FSV LaTeX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Cala, Petr: *Ability bias in the returns to schooling: How large it is and why it matters*. Diploma Thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 106. Advisor: doc. PhDr. Zuzana Havránková, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

BMA Bayesian Model Averaging

EK Endogenous Kink

FMA Frequentist Model Averaging

FAT-PET Funnel Asymmetry Tests - Precision Effect Testing

GLS Generalized Least Squares

HCEF Human Capital Earnings Function

IV Instrumental Variable

MAIVE Meta-Analysis Instrumental Variable Estimator

OLS Ordinary Least Squares

PMP Posterior Model Probability

PIP Posterior Inclusion Probability

RoBMA Robust Bayesian Model Averaging

WAAP Weighted Average of Adequately Powered

WLS Weighted Least Squares

Chapter 1

Introduction

Text

Chapter 2

Private returns to education

2.1 Human capital theory and Mincer equation

The backbone of research into the returns to schooling topic lies in the Human Capital Theory, conceptualized first by Becker (1962). The idea is simple - investments in education should improve one's productivity, resulting in increased income over time. The author bases the calculation on a simple cost-benefit relationship; if an individual puts time, effort, and money into their education, this investment should bring returns later on in the form of increased earnings. Schultz (1961) then argue that the crucial factor behind the increase in earnings is the heightened productivity of the individual gained during the years spent in school.

Roughly a decade later, Mincer (1974) proposed a vital extension to this theory, quantifying this relationship in a model called the Human Capital Earnings Function (HCEF). In this equation, usually referred to as the *Mincer equation*, the log of one's earnings can be expressed as an additive function of a linear education term and a quadratic experience term. Rigorously, we can write this semi-logarithmic relationship as

$$ln(Y_i) = \alpha + \beta S_i + \gamma_1 X_i + \gamma_2 X_i^2 + \epsilon_i, \tag{2.1}$$

where $ln(Y_i)$ denotes the log of earnings of an individual i, S_i represents their attained years of schooling, X_i stands for the years of work experience of said individual, and ϵ_i captures the individual-specific error. In cases where

the individual's experience is absent, Mincer (1974) proposes using a measure of potential experience instead. This can be calculated as

$$X_i = A_i - S_i - 6, \tag{2.2}$$

where X_i denotes the potential experience, A_i represents the individual's age, and S_i stands for the completed years of schooling. Six is a constant; we assume that the individual starts their education at age six.

Over the decades, this equation has been subjected to the scrutiny of many research papers (Ashenfelter & Krueger 1994; Heckman et al. 2006; Card & Krueger 1992); naturally, this scrutiny raised several questions regarding the functional form of the equation. Heckman et al. (2003) subject the equation to a thorough analysis by relaxing the proposed functional form, and arrive at results that differ substantially from the ones drawn from the Mincer equation. As for specific extensions to the existing equation, Card (1999) proposed adding control variables to the Mincer equation, including race, geographic region, and union membership. Using these new controls, they highlighted the importance of an individual's location factors and their role in determining one's income. Psacharopoulos & Patrinos (2004) highlighted the importance of the individual's socioeconomic background as a predictor for earnings with findings that firmly back up their claim. Belzil & Hansen (2004) then extend the equation to account for individual heterogeneity by employing a dynamic programming model of schooling decisions.

Among the many available methods for estimating this relationship, OLS is the most common approach (Ashenfelter et al. 1999; Card 1999). However, OLS estimates suffer from several estimation problems, including sample selectivity, omitted variable bias, and measurement error bias, as noted by Aslam (2007), among others. Equations using year cohorts (Angrist & Pischke 2009), Heckman's correction for sample selectivity (Heckman 1979), or fixed-effects are among the several that tackle these issues.

Still, there exists one other important issue in the literature that plenty of authors choose to avoid, one I firmly believe should be addressed - the issue of unobserved ability.

2.2 Ability bias in the Mincer regression

One variable stands out among the many that could be included in the Mincer regression, and the omission of which may lead to biased estimates - individual ability. There exists a plethora of research in psychology to show that general intelligence is one of the most reliable predictors of one's success (Gottfredson 1997; Deary et al. 2007; Deary 2020; Ozawa et al. 2022). This measure can then be quantified; researchers refer to it as the g-factor. When this factor is included in the regression, other determinants of an individual's life outcomes suddenly lose their predictive power, coining the phrase *"not much more than g"* (Ree et al. 1994). Heckman & Rubinstein (2001) support this claim by examining the role of non-cognitive abilities in determining earnings and educational attainment, finding out that they serve as crucial predictors in these areas of economic development.

Regarding policy making, the predictive analysis prevalent in psychology helps us only a little (Almlund et al. 2011). While highly useful when placing an individual into the labor market, predictive analysis deals with correlations rather than causal effects, which are the focus of policy analysis. Indeed, without a way to assess the impact of the policy changes, evaluating the quality of said change is impossible. Undoubtedly, one of the major objectives of education policies lies in the improvement of one's capacity to succeed in the labor market. However, if the estimate of the returns to education is biased, these policies could quickly be rendered inefficient and misguided.

Herrnstein & Murray (2010) bring these two issues together in a study that reveals how economic returns tend to rise with higher individual ability. Bowles et al. (2001) provide more evidence by showing that the returns to schooling in the Mincer equation tend to be inflated when ability (or other measure of cognitive performance) is omitted. Over the years, the term *ability bias* that describes this phenomenon has been subjected to the scrutiny of research (Heckman & Vytlacil 2001). Multiple researchers attribute little to no importance to this issue (Ashenfelter & Rouse 1999). Apart from suggestions for its omission (Blackburn & Neumark 1993), some claim that non-cognitive abilities hold no less predicting weight (Heckman & Rubinstein 2001). Griliches (1977), for example, finds out that the bias is either small or negative, and Patrinos (2016) argues that adding more variables to the equation will not solve the problem; instead, it may introduce new biases on its own.

A whole new branch of research into ability bias lies within natural experi-

ments. Some economists (Ashenfelter & Krueger 1994; Berman et al. 2003), for example, turned to twin studies to identify the role of education, as other factors (such as socioeconomic background, abilities, preferences, etc.) are nearly identical with twins. I must, however, address two points of criticism prevalent in the literature (Kenayathulla 2013). Firstly, there is no way to guarantee the exogeneity of ability. In other words, if ability would have both an individual and a family component, the latter would be endogenous to schooling, failing to a potentially still biased estimate. Secondly, measurement errors pose a particular threat to the result validity, as those errors could explain most twin-level differences across the population (Ashenfelter et al. 1999). Nonetheless, twin studies provide an intriguing alternative way to survey the ability bias issue from another perspective, although most authors overlook this possibility entirely.

On balance, the ability bias issue lives in a niche spot of researchers' consciousness. On top of the lack of consensus on the theoretical side of the research, the practical side is just as discordant. A growing practice has had researchers choosing a proxy in their estimation to control for ability indirectly, usually with parental education, marital status, or distance to school, among others (Blundell et al. 2001). The authors often acknowledge that their estimates could be plagued by this bias but fail to obtain the data necessary for its treatment (Agrawal 2012; De Brauw & Rozelle 2008). Other times, the issue gets overlooked entirely, and the authors focus either on the simplest or a more complex form of the Mincer regression (Angrist 1995; Sinning 2017). Given the disunified practice, I proceed to answer the following questions. Does this ability bias matter? How large is it? If we control for this bias, how do the returns to education change?

2.3 Existing research

Before answering the questions, it is crucial to look at and acknowledge the existing meta-analyses that have already tackled these issues before me. As of me writing this paper and to the best of my knowledge, these are all of the meta-analyses that have been conducted on the topic of returns to education thus far - Psacharopoulos (1994); Fleisher et al. (2005); Churchill & Mishra (2018); Psacharopoulos & Patrinos (2018); Patrinos & Psacharopoulos (2020); Cui & Martins (2021); Iwasaki & Ma (2021); Ma & Iwasaki (2021); Wincenciak

Table 2.1: Existing meta-analyses choose to tackle different issues

Study name	**AB**	**AB***	**PB**	**PB***	**Method**
Psacharopoulos (1994)	.	.	.	.	.
Fleisher et al. (2005)	.	.	.	.	✓
Churchill & Mishra (2018)	.	.	✓	✓	✓
Psacharopoulos & Patrinos (2018)	.	.	.	.	.
Patrinos & Psacharopoulos (2020)	.	.	.	.	.
Cui & Martins (2021)	.	.	✓	✓	✓
Iwasaki & Ma (2021)	.	.	✓	.	✓
Ma & Iwasaki (2021)	.	.	✓	✓	✓
Wincenciak et al. (2022)	✓	✓	.	.	✓
Horie & Iwasaki (2023)	.	.	✓	.	.
Number of studies:	1	1	5	3	6
Percentage of studies:	10%	10%	50%	30%	60%

Note: This table lists (to my knowledge) all existing meta-analyses on the topic of returns to education, along with information about methodology each of them chooses to employ. A check-mark means the study does tackle the corresponding issue. The last two rows display the number of studies dealing with each issue in absolute and relative terms. AB = The study analyses ability bias as a predictor for returns to schooling, AB* = The study finds that ability bias is a strong predictor for returns to schooling, PB = The study addresses publication bias, PB* = The study finds publication bias in its data, Method = The study addresses the type of methodology used by the examined studies.

et al. (2022); Horie & Iwasaki (2023). In Table 2.1, I outlined how each of these studies tackles the several main points of existing research.

Out of these ten studies, only the paper by Wincenciak et al. (2022) attempts to directly answer the role of ability in estimating returns to schooling. They find that ability is a significant predictor of returns to education (about 0.8-0.9% points) when controlled for. They conclude that the omission of ability bias may lead to biased estimates of the discussed effect. As for other studies, Fleisher et al. (2005) and Patrinos & Psacharopoulos (2020) acknowledge the presence of ability as a potential predictor in the Mincer regression but either dismiss its validity or choose not to analyze the issue in depth.

Five studies then deal in any form with publication bias (for brevity, I will not list them; refer to Table 2.1 for detail). Three of these studies then find a presence of publication bias in the literature, while the other two do not.

Lastly, six of the ten existing meta-analyses include control in any form for methodology in their approach. Mainly, this involves putting a single control such as Instrumental Variable (IV) or Ordinary Least Squares (OLS) into their models. None of the studies then compare more methods to each other.

Indeed, no single study exists that would bring all these issues together and try to answer all of them. This, together with other vital points, should be the main focus of this thesis, as explained in the following section.

2.4 My contribution

What I hope to bring into the field with this thesis can be summarized in the following way.

First, as outlined in Section 2.3, only one meta-analysis on the role of ability bias in returns to education exists thus far (on top, this paper has been published only after conceptualizing this thesis). Although I may not be the first to consider ability bias as a significant predictor of the effect of education, it is far from feasible to claim that the ability bias issue has been explored - far from it. I hope to thoroughly examine how ability plays its part as a predictor of returns to schooling, observe whether it is statistically and economically significant and whether it should be treated for. Furthermore, the existence of a meta-analysis on the topic means that I can now compare my results with the existing ones, which should ultimately bring more credibility to the issue overall.

Second, by clearing up the uncertainty regarding the influence of ability bias on one's future income, I can suggest more efficient ways to indirectly control for ability or even highlight the importance of obtaining data through which the researchers can control for this bias. Given the existing heterogeneity in the current research (especially regarding ability bias), this may help guide the authors in their estimation strategies and finally contribute to the quality of research findings in the future.

Third, I hope to identify the individual effects that different estimation methods may systematically have on returns to education. Even though over half of the existing meta-analyses address this issue, none directly compares all of the available methodologies within the literature. Given that the dataset I will assemble and use to test for this relies primarily on a search query for the choice of studies, the literature set should provide the most representative form of the existing literature possible and capture nearly all methods used in practice.

Fourth, I will focus thoroughly on the issue of publication bias to find systematic misuse of result reporting. By employing the most modern state-of-the-art methodology such as the MAIVE estimator (Irsova et al. 2023) or Robust Bayesian Model Averaging (Maier et al. 2022) in addition to the battery of the standard FAT-PEESE-PET tests and more, I attempt to bring the most robust results out of all existing analyses thus far. Looking at the results of the five that have tried to answer the issue, no consensus exists here either (three

claim the presence of publication bias, while two argue for the lack thereof). More scrutiny should help clear out the uncertainty about publication bias and provide even more robustness to the results.

Fifth, I will look at the role of individual variables regarding the effect behavior using novel technology such as Bayesian and Frequentist Model Averaging. Which variables are the most influential drivers of the returns to education effect? What is their economic significance? What would be the true effect if a best-practice effect could be derived from the literature that would correct for the aforementioned detected biases? None of the existing research tackles any of these questions, and I hope this approach will contribute to their clarification.

Sixth, I will construct an entirely new dataset including only natural experiments conducted on twins (so-called *twin studies*) and rerun the analysis using this dataset. By removing the differences in socioeconomic factors that usually exist in the subject sample, this approach should serve as a robustness check to more precisely identify education's role in affecting the twins' future earnings. To keep things concise and not branch off too far, I intend to skip (or at most, gloss) over the results regarding publication bias, heterogeneity, and best-practice estimate. Instead, I shall focus on how ability bias changes with this new twin study dataset. In any case, this should help me further validate the robustness of my results.

Next, I present several technical extensions as an improvement to the code quality of the analysis. As the first one, I provide R code for the Endogenous Kink method by (Bom & Rachinger 2019). So far, to the best of my knowledge, the code for this method is publicly available only in the STATA software. I hope to facilitate research to a potentially sizable pool of researchers who do not work with or hold the license to STATA by providing the code for said method purely in the programming language R. Several validity checks are also included in the new code to make sure it runs smoothly and without hiccups. Albeit a trifling task, I believe it will aid further researchers shine a brighter light on their results.

As the second technical extension, I upgrade the existing code of the STEM method (Furukawa 2019) to work up to orders of magnitude times faster than the available source code.[1]

As the last extension, I provide an all-encompassing R code in the form

[1]Tested on the full master dataset of length 1754, the improvement cuts down the source code run time of 99.52 seconds to only 2.84 seconds, averaged over ten runs.

of several scripts that can be used together to replicate the whole analysis without effort.[2] With over 7000 lines of code, the project allows the user to run, see, and customize every method from a single point of entry. All results are automatically exported and saved in a single, small-sized, and easily distributable *.zip* file. With best-practice methods from software engineering, including tests, validation checks, a cache system, and much more, anyone can now access complex meta-analysis methods and run them all in seconds.

[2]Available at https://github.com/PetrCala/Diploma-Thesis.

Chapter 3

Data

3.1 Literature search

I employ the Google Scholar search engine with its full-text search capabilities to assemble my dataset. A query constructed using a combination of keywords helps me narrow down the results into studies dealing with ability bias, private returns, and education. After several modifications to ensure the query generates consistent results within the scope of interest, I obtained the query's final form. To view this form, refer to Appendix A.

I ran the definitive search on January 23, 2023, and received 574 hits. To achieve absolute consistency, I employed web scraping and automatic data preprocessing tools and denoted vital information about all 574 studies during a single day. These included the authors' names, publication information, the number of citations, and the impact factor of the journal the study was published in[1]. To avoid duplicate results and guarantee the uniqueness of each hit, I also extracted the study result IDs.

I then went through the first 200 studies and identified 78 as eligible for data collection. This means they report an estimate of a regression of wage on a schooling variable. Despite a sizable number of studies being eliminated due to lack of data, 129 of these studies (corresponding to over 60% of the surveyed sample) were at least relevant to the topic, validating thus the quality of the query.

In accordance with the reporting guidelines for meta-analysis by Havránek et al. (2020), I define the following criteria that will help me narrow down the

[1]In case of an unpublished study, I set the impact factor to 0.

study list into its final form. For a study to be included in the dataset, it must fulfill several criteria.

First, the study must report one or more estimates from an equation of any form of wage on a schooling variable (years of schooling or completed level of education), along with their standard errors or corresponding t-statistics. Without the last two mentioned, there would be no way to compare the strength of the effects. Furthermore, there must either be a traceable statistic associated with every estimate that signifies the number of degrees of freedom or sample size from the regression or, in case neither of these is provided, there must be a number denoting the number of subjects for the experiment. In such cases, it must be evident that the sample size corresponds to the reported estimates. Due to the thoroughness of the initial screening, only four studies of the 78 did not fulfill these criteria and were thus removed. This leaves a total of 74 studies eligible for collection.

To retain as much information about the research field as possible, I choose *not* to discard studies of varying quality, including unpublished papers, graduate theses, dissertations, etc. There is no consensus in the existing literature on which approach should be taken, as highlighted by Stanley (2001). Even though the author advises careful consideration when including unpublished studies, they also acknowledge that their omission could create a new publication bias instead. The inclusion approach is also supported by Cook et al. (1993), who found that numerous meta-analysis researchers and methodologists believe data from unpublished studies should not be discarded if one aims to synthesize the available information objectively.

However, upon closer inspection of the initially generated list of studies, I observed that several highly influential studies from the field were missing, such as those by Angrist & Krueger (1991), Staiger & Stock (1997), or Heckman et al. (2006). These failed to get identified as relevant by the query and did not appear in the search results. To ensure the whole field of relevant literature is encompassed, I employ the snowballing method to incorporate these crucial studies.

The use of the snowballing method itself is debatable, and an argument can be made for its avoidance, as the data search suddenly becomes hard to replicate. Indeed, having only one search query would be ideal, but the unfortunate lack of several highly relevant studies seemed a reason enough to give snowballing a green flag. Given that several meta-analyses (Psacharopoulos 1994; Fleisher et al. 2005; Psacharopoulos & Patrinos 2018) and highly cited stud-

ies (Card 1995; Heckman et al. 2006; Psacharopoulos & Patrinos 2018) have presented results that have since been many times reviewed and are indeed well-established, the omission of some of these crucial studies seems detrimental to the quality of the analysis.

With the decision to add these studies, I conducted a meticulous search of bibliography references from the already identified studies. This search then yielded 55 additional papers that significantly contribute to the topic. After applying the earlier-mentioned criteria, I narrowed this list to 41 highly relevant and collectible papers. Combined with the 74 studies identified during the query search, the final list consists of 115 studies, which I will refer to these as the *primary studies*. These studies should thoroughly encapsulate the existing literature's findings and methodologies and provide a more robust representation than the query search subset. The final list of studies can be found in Appendix A, together with a PRISMA flow diagram summarizing the literature search.

3.2 Interpreting of the Effect in Question

A glance into the assembled literature set reveals an important issue I must address before explaining the data collection process. That is, what is the effect that we are collecting?

Many studies in the set (Sackey 2008; Leigh 2008; Bartolj et al. 2013) use schooling in levels rather than years. The most prominent argument for this choice is undoubtedly the lack of data on the exact years of education. Further, this approach is certainly a valid way of estimating the Mincer equation, as one can observe how different levels of educational attainment contribute to the log of an individual's earnings. Quantitatively, we can extend the Equation 2.1 to the following form:

$$ln(Y_i) = \alpha + \beta_1 PRIM_i + \beta_2 SEC_i + \beta_3 HIGHER_i + \gamma_1 X_i + \gamma_2 X_i^2 + \epsilon_i, \quad (3.1)$$

where $PRIM$, SEC, and $HIGHER$ represent dummy variables for primary, secondary, and higher education, respectively. The rest of the variables and their explanation is the same as in Equation 2.1. Note that the levels included in the regression do not necessarily have to conform to the three dummies outlined here; quite the contrary. In practice, the authors (Gill & Leigh

(2000) as an example) choose schooling levels that best represent their data. This includes adding in variables representing attainment of a Bachelor's degree, Master's degree, or even country-specific education levels.

The critical question is, having these different levels of schooling, can you calculate the returns to an additional year of schooling for all these level coefficients so that the estimates are directly comparable? The answer is yes, you can. When comparing returns of one schooling level to another, variations of the following formula (such as in Agrawal (2012)) can be used to quantify the relationship between schooling in levels and years of schooling:

$$S_i = (1 + \beta_{i,higher} - \beta_{i,lower})^{\frac{1}{Y_{i,higher} - Y_{i,lower}}} - 1, \tag{3.2}$$

where S_i denotes the effect an additional year of schooling has on the log wage of an estimate i, $\beta_{i,higher}$ and $\beta_{i,lower}$ are the coefficients from the Mincer regression associated with the higher and lower schooling levels respectively. Finally, $Y_{i,higher}$ and $Y_{i,lower}$ are the years it takes to complete the higher and lower schooling level, respectively.

This form of the equation assumes there are two levels of schooling present in the regression, and its result is the return to a year of schooling within these two (i.e., when comparing primary to secondary schools, the resulting coefficient would denote how much each year of secondary school contributes to an individual's earnings). Suppose no other level is available for comparison, such as when calculating the returns to schooling for the first level coefficient in the equation. Then we can plug 0 for the other schooling level's coefficients, which reduces the equation to the following form:

$$S_i = (1 + \beta_i)^{\frac{1}{Y_i}} - 1. \tag{3.3}$$

Here, β_i is the Mincer regression coefficient associated with the attained schooling level of an estimate i, and Y_i denotes the years required to obtain said education level.

After transforming the effect, one must also handle the standard errors and resulting t-statistics. Given that the standard error does not directly carry through nonlinear transformations (which both Equation 3.2 and Equation 3.3 are), it is necessary to derive the standard error in another way. For this, I use the *delta method* (Ziegel 2002), which helps me calculate the standard error. I run most of the calculations using the R *deltamethod* function (Fox & Weisberg 2018), where only the functional form is required along with the respective

coefficients. I calculate the t-statistics only after obtaining the transformed estimates and their standard errors; this ensures the validity of publication bias methods used later in the work. Further, I scale all the numbers by a factor of 100 for direct interpretability of the effect as a percentage return to an additional return of schooling.

The last important question to answer is whether, after unifying the different types of the effect, there would not appear any kind of systematic pattern in the literature that could invalidate the results. Indeed, in the meta-analysis of Churchill & Mishra (2018), the authors use the FAT-PET-PEESE tests to (among other things) study whether the reported returns to an additional year of schooling vary systematically depending on the education type measure. They find that studies using years of schooling report higher estimates than those using education levels. Given this finding, I choose to include a variable in my dataset that controls for the type of estimate reporting used. In theory, such a coefficient should be 0 (meaning there is no systematic difference between reporting in years and levels). Chapter 5 reveals whether whether this holds.

3.3 Dataset assembly

Having the effect interpretation cleared up, I proceed to data collection. From 115 relevant studies, I collected 1754 estimates of the effect together with dozens of other variables that helped me capture heterogeneity within the literature. Apart from the necessary numeric statistics such as standard error, t-statistic, or degrees of freedom, I also collected over 40 variables categorizing the effect type, study characteristics, spatial/structural variation, estimation method, and publication characteristics. See Table 5.1 for a complete list of these variables. The table also contains descriptions and summary statistics of the variables.

Upon closer inspection, I observed that studies in my dataset can be split into four categories based on their approach toward ability. I capture this in the variable *Ability*, where the categories can be defined as follows:

- *Ability: Direct* - The study directly includes a measure of ability in the regression. This can mean a score from an IQ test, a measure of language ability, or any other kind of ability. Grogger & Eide (1995) or Van Praag et al. (2013) are good examples of this approach.

- *Ability: Proxy* - The authors use a proxy for ability instead, such as a relative's education level or the number of siblings. Often, this is associated with the use of Instrumental Variable regression. Card (1995) or De Brauw & Rozelle (2008) use such proxies.

- *Ability: Uncontrolled* - The authors address the issue of ability bias in their work but can not or choose not to add any measure or proxy for ability into the regression. This could be due to a lack of data or their reasoning for the inconsequentiality of ability bias. (Angrist & Krueger 1991; Fang et al. 2012)

- *Ability: Unmentioned* - There is no mention of ability or ability bias anywhere in the study. The results are typically reported in the form of a simple Mincer regression. Staiger & Stock (1997) or Acemoglu & Angrist (1999) fall into this category.

As far as the other variables are concerned, I was, in most cases, able to collect all the necessary data. However, some variable groups still had to be dropped for the lack thereof. Topics such as education field (STEM, Medicine, Law,...), regression type (Mincer vs. Discounting), or school type (Private vs. Public) were all addressed within only a few, if any, studies, making them infeasible to collect. On the flip side, I identified and added a handful of variables I had not intended to add initially, such as marriage control or residential area type (rural vs. urban). I also added data on the country-year-specific level (meaning it differs for each country-year pair), such as minimum wage or median household expenditure. I also added a variable on the country-year level capturing the Academic Freedom index, the data for which I obtained from the dataset by Coppedge et al. (2023).

Regarding study-specific variables, such as the number of citations, publication status, or impact factor, I ensured that all these could be directly comparable by measuring them in a single day - January 23, 2023. Any changes within these variables for the included studies after this date are not considered.

Further, I can use the human capital earnings function described in Chapter 2 and take the potential experience measure from Equation 2.2. Using this relationship, it is possible to derive missing values of either one of the mean years of schooling, the mean years of experience, or the mean age, provided the other two statistics are reported. For example, if a study includes the subjects' mean age and mean years of schooling but omits the mean experience, it can

be calculated as *age - schooling - 6*. On the flip side, there are times when a study fails to report at least two of these variables. In those cases, I leave the underivable values empty.

However, methods used later in this work require the absence of missing observations in the data. To treat this, I use clever interpolation to fill in the missing observations to copy the existing information as closely as possible. For variables of *float* type, such as minimum wage, age of subjects, or freedom index, and variables of *dummy* type, such as wage earners vs. self-employed, I use the median of the existing data for the given variable. At other times, the variable can be aggregated at the country level. In that case, the interpolation happens at the same level, meaning that the medians are calculated for individual countries, not across the whole dataset. For percentage variables (such as the ratio of subjects living in urban vs. rural areas), the mean of the data is used, aggregated again on a country-specific level. This ensures that the ratios always sum up to one and simultaneously capture the situation representing the study's environment as closely as possible.

With these transformations, I obtained the final form of the dataset with 1754 observations and more than 150,000 data points in total. To see the data frame, refer to the files appended with this work[2]. Alternatively, you can also find the data set on the project's GitHub page.

3.4 Initial analysis

After cleaning the dataset and double-checking that all calculations were correct, I checked the effect behavior through various subsets of data. In Table 3.1, you can find the summary statistics of the effect under these subsets, while Figure 3.1 offers a graphical insight instead. When splitting the data into subsets where it was unclear what point to choose for the split (such as the case of *Observations*, *Data Year*, or *Citations*), I used the median of the variable in question. At other times, such as for variables reported in ratios, I used 0.5 (50% of observations) as the split point.

First, I quickly address the numeric results. As a baseline for the rest of the work, we can observe that the unweighted mean of the effect across all data equals 7.476 (7.652 for the data weighted by the number of estimates reported

[2]The missing values are interpolated only upon the script run and not within the raw data. Running the code will also inform you about the missing values, their handling, etc. Find the code within the appended files, also.

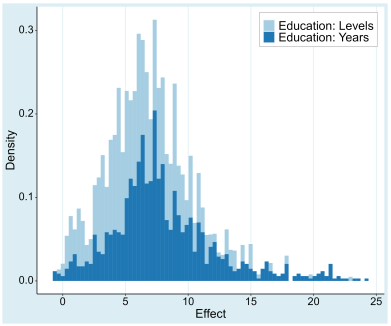
Table 3.1: Mean statistics across various subsets of data

	Unweighted			Weighted			
	Mean	95% conf. int.		Mean	95% conf. int.		N. obs
All Data	7.476	-1.224	16.176	7.674	-1.026	16.374	1,754
Estimate characteristics							
Estimate: City	8.579	-2.201	19.359	7.674	-3.106	18.454	208
Estimate: Sub-region	7.025	-2.334	16.384	7.674	-1.685	17.033	174
Estimate: Region	7.231	-1.079	15.541	7.590	-0.720	15.900	542
Estimate: Country	7.478	-0.505	15.461	7.664	-0.319	15.647	692
Estimate: Continent	7.331	-1.565	16.227	7.999	-0.897	16.895	138
Observations >= 6476	7.150	-0.392	14.692	7.524	-0.018	15.066	882
Observations < 6476	7.806	-1.888	17.500	7.832	-1.862	17.526	872
Data Characteristics							
Study Size >= 20	7.187	-1.766	16.140	6.928	-2.025	15.881	884
Study Size < 20	7.769	-0.631	16.169	7.938	-0.462	16.338	870
Yrs. of Schooling >= 10.9	7.692	-0.358	15.742	7.793	-0.257	15.843	881
Yrs. of Schooling < 10.9	7.257	-2.039	16.553	7.562	-1.734	16.858	873
Yrs. of Experience >= 19.48	7.595	-0.564	15.754	8.104	-0.055	16.263	901
Yrs. of Experience < 19.48	7.350	-1.885	16.585	7.243	-1.992	16.478	853
Cross-sectional Data	7.559	-0.990	16.108	7.520	-1.029	16.069	634
Panel Data	7.429	-1.358	16.216	7.771	-1.016	16.558	1,120
Data Year >= 1999	8.214	-1.349	17.777	8.276	-1.287	17.839	901
Data Year < 1999	6.696	-0.693	14.085	7.144	-0.245	14.533	853
Spatial/structural variation							
Higher Education >= 0.5	8.599	2.556	14.642	9.035	2.992	15.078	311
Higher Education < 0.5	7.234	-1.870	16.338	7.414	-1.690	16.518	1,443
Wage Earners >= 0.5	7.523	-1.234	16.280	7.731	-1.026	16.488	1,632
Self-employed > 0.5	6.848	-0.986	14.682	6.846	-0.988	14.680	122
Male >= 0.5	7.180	-1.440	15.800	7.450	-1.170	16.070	1,298
Female > 0.5	8.318	-0.406	17.042	8.439	-0.285	17.163	456
Private Sector >= 0.5	7.628	-1.186	16.442	7.772	-1.042	16.586	1,540
Public Sector > 0.5	6.377	-1.126	13.880	7.022	-0.481	14.525	214
Rural >= 0.5	7.080	-3.255	17.415	7.388	-2.947	17.723	176
Urban > 0.5	7.520	-0.978	16.018	7.712	-0.786	16.210	1,578
High Income Countries	7.023	-0.260	14.306	7.141	-0.142	14.424	889
Middle Income Countries	7.868	-1.914	17.650	8.035	-1.747	17.817	761
Low Income Countries	8.476	-1.994	18.946	9.716	-0.754	20.186	104
Mean Age >= 37	7.570	-0.380	15.520	8.180	0.230	16.130	900
Mean Age < 37	7.376	-2.051	16.803	7.142	-2.285	16.569	854
Estimation method							
Ability: Direct	6.233	-0.419	12.885	6.417	-0.235	13.069	236
Ability: Proxied	8.906	-2.705	20.517	9.040	-2.571	20.651	357
Ability: Uncontrolled	7.675	-0.529	15.879	7.619	-0.585	15.823	745
Ability: Unmentioned	6.604	-0.211	13.419	7.106	0.291	13.921	392
Control: Age	8.320	-1.202	17.842	8.598	-0.924	18.120	604
Control: Age^2	9.094	-0.039	18.227	9.296	0.163	18.429	482
Control: Experience	7.002	-1.385	15.389	7.130	-1.257	15.517	1,064
Control: Experience^2	7.177	-1.396	15.750	7.139	-1.434	15.712	898
Publication characteristics							
Impact Factor >= 0.191	7.021	-0.874	14.916	7.338	-0.557	15.233	877
Impact Factor < 0.191	7.930	-1.427	17.287	8.068	-1.289	17.425	877
Citations >= 80	7.178	-0.826	15.182	7.531	-0.473	15.535	892
Citations < 80	7.784	-1.547	17.115	7.815	-1.516	17.146	862
Study: Published	7.222	-0.739	15.183	7.654	-0.307	15.615	1,340
Study: Unpublished	8.298	-2.300	18.896	7.758	-2.840	18.356	414

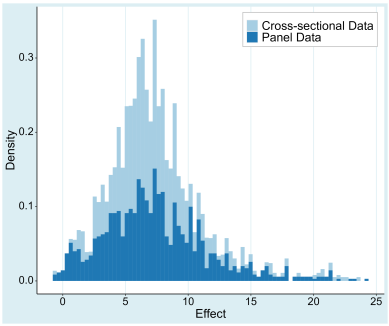
Note: This table presents basic summary statistics of the returns to an additional year of schooling coefficient calculated on various subsets of the data. Unweighted = Original dataset is used. Weighted = Estimates are weighted by the inverse number of estimates reported by each study. OLS = Ordinary Least Squares. For cutoff points, medians are used except for dummy variables, where the cutoffs are 0.5.

Figure 3.1: Graphically observing the effect across subsets of data

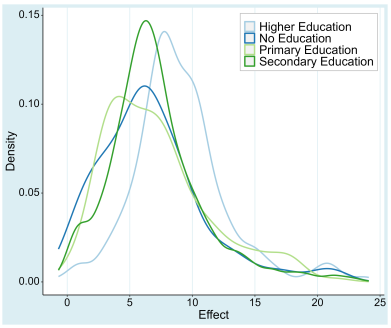
(a) Education type



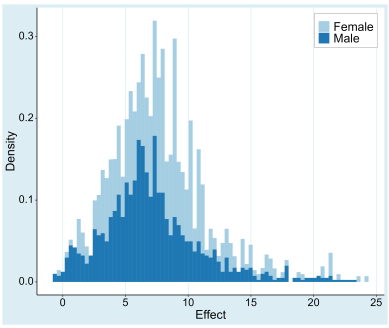
(b) Data type



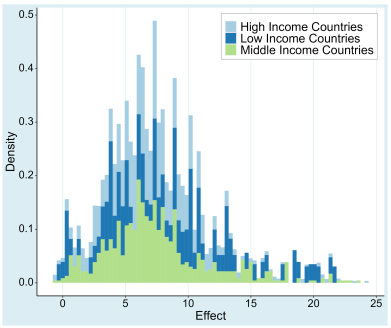
(c) Highest education



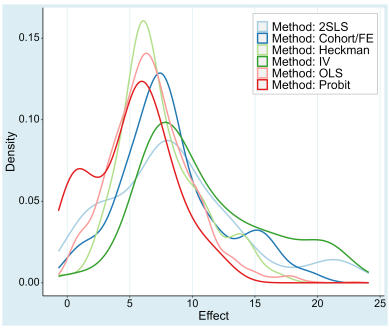
(d) Gender



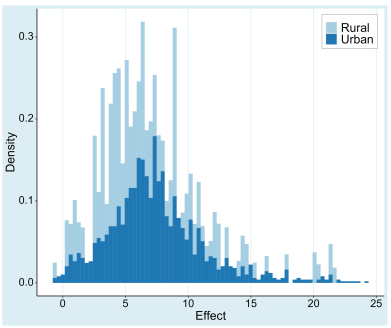
(e) Country wealth



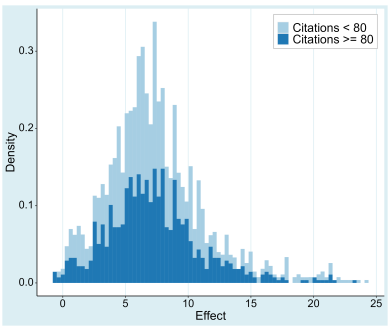
(f) Estimation method



(g) Ability



(h) Citations



Note: This figure displays histograms and density lines for different subsets of data, where the effect of an additional year of schooling on returns is displayed on the x-axis against its density on the y-axis. For Figure 3.1h, the data median is used to determine the subsets. For a description of variables used in this figure, see Table 5.1.

per study). As a reminder, this can be interpreted as a 7.456% increase in log wage per additional year of attained schooling and falls well into the expected range, comparing this estimate to the results of other works. As such, this brief insight can serve as a sanity check that there is nothing immediately wrong with the data collection. When comparing to individual studies, this mean is slightly lower than Psacharopoulos & Patrinos (2018) who claim about 9% average returns to schooling, but a bit higher than the findings of Fleisher et al. (2005) who report returns between 5 and 6 percent on average. My results also align well with the only study dealing in detail with ability bias, Wincenciak et al. (2022), where the authors also report a 7% figure for the average effect. Note that the suggested figure of 7.476 does not account for publication bias and should thus be treated only as a benchmark for further comparison.

Concerning other subsets of data, there appears to be variety in several variable categories, including the age of data, economic status of countries, study publication status, or, perhaps more interestingly, ability. Estimates aggregated on the city level can be associated with higher estimates of the effect (8.579), yet this difference disappears entirely after accounting for the study size (7.674). The same is true for estimates from unpublished studies (8.298 vs. 7.758). On the other hand, estimates associated with other variables remain higher than their counterparts, even through weighting. These include smaller sample size estimates, smaller studies, newer data, estimates for subjects with higher education, countries with low income, female subjects, or studies with a smaller impact factor. Perhaps most interestingly, the mean estimate is also higher for studies that proxy for ability and marginally for those that do not control for it.

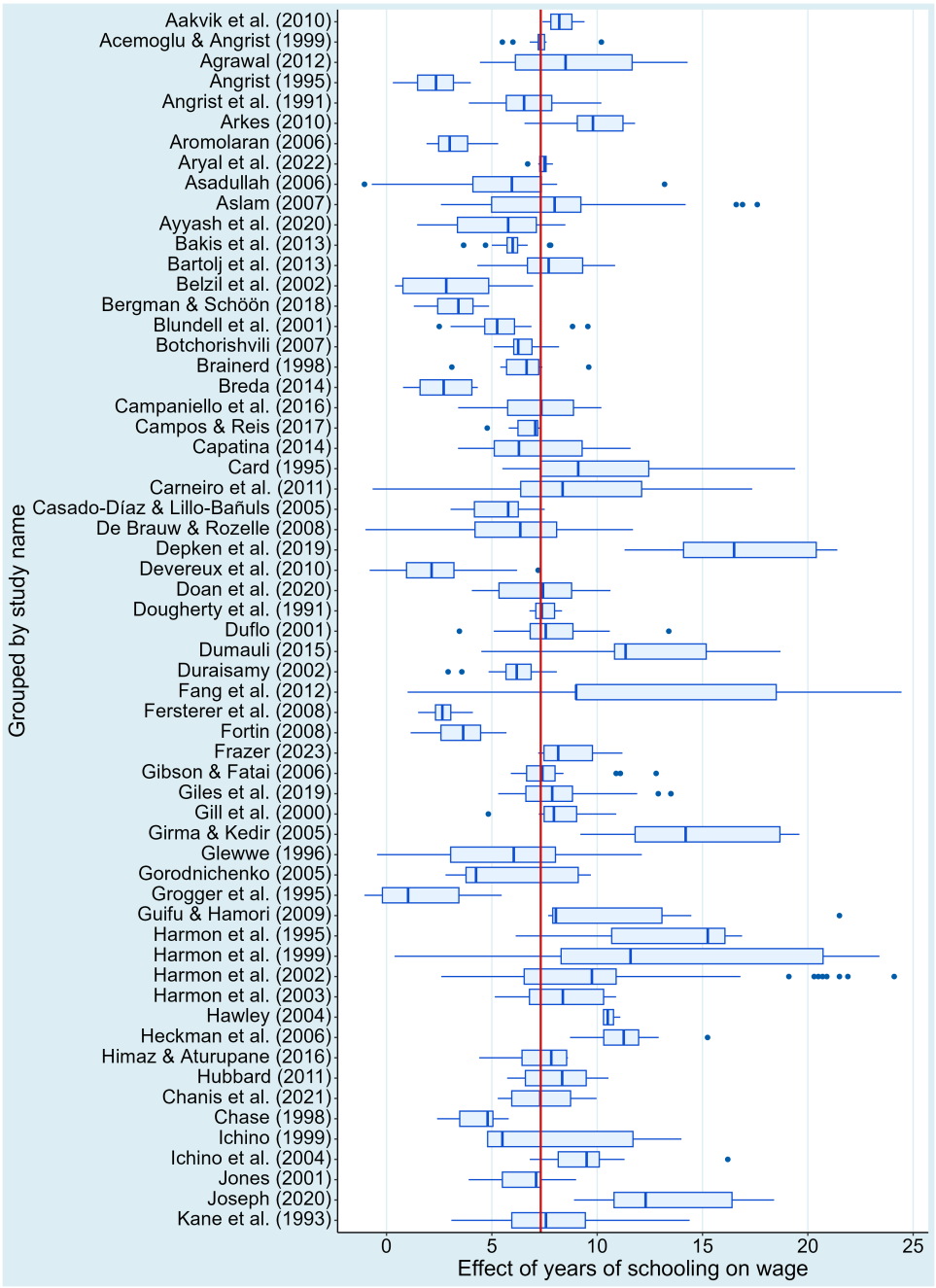
However, given the wide confidence intervals associated with all these subsets, one should take all these claims with a grain of salt. Furthermore, these differences are only marginal and should not serve as concrete evidence of a clear trend. Moreover, the mean could hardly be considered a statistical measure with perfect information; more data scrutiny will surely be necessary. This I will focus on in Chapter 5.

As for the graphical insights that can be obtained from the data, Figure 3.1c confirms that the distribution of estimates associated with higher education holds perhaps estimates of higher returns than its counterparts. Further, the right tail of the *Ability: Proxied* variable distribution appears the heaviest out of the four subcategories. This may suggest that including a proxy for ability often yields higher estimates in ranges that other approaches seldom report.

The right-side distribution tails also vary notably for the *Method* variable, with IV regression approach reporting the highest estimates of all techniques. In contrast, methods such as Probit or OLS sporadically yield coefficients of unusually high value.

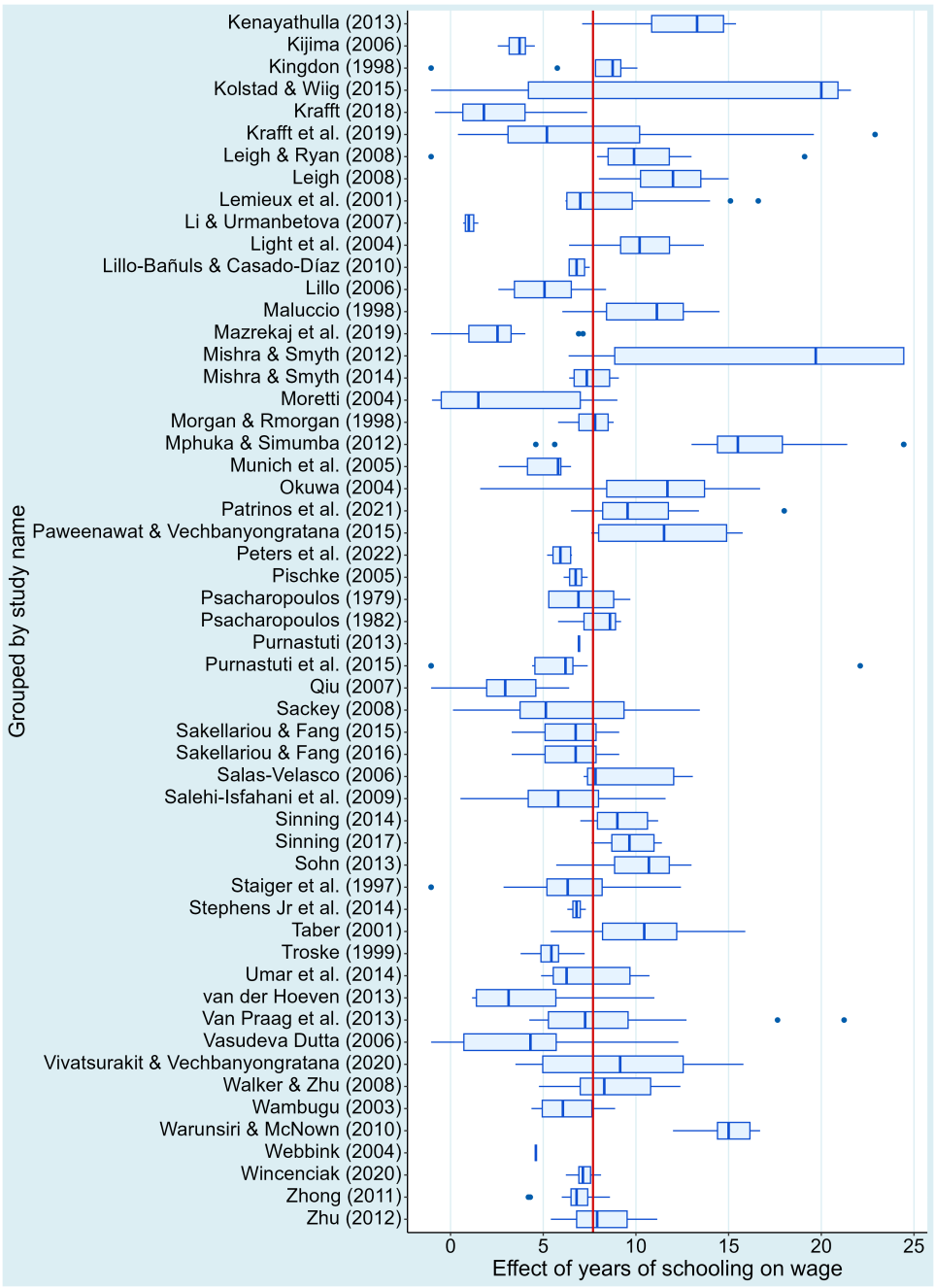
To highlight the differences between individual studies, I also include box plot of study-level clustered data in figures 3.2 and 3.3 (in the Appendix A, you may also find a country-level box plot for additional insight into the data). For clarity of presentation, I present two plots instead of one due to the large number of studies within the dataset. The split is done arbitrarily after 60 studies, ordered alphabetically. Despite the evident presence of outliers in some cases (Asadullah 2006; Harmon et al. 2002), the studies, in most cases, report results close to the mean; only a handful of studies stand in the plot far out from the mean line. Studies of Depken et al. (2019); Girma & Kedir (2005), or Mphuka & Simumba (2012) report peculiarly high estimates, while studies like Angrist (1995); Li & Urmanbetova (2007), or Webbink (2004) never report an estimate above 5% according to my calculations. To detect and amend for any potential miscalculations and human error, I double-checked the source data along with the calculations and, after this validation, proclaimed the dataset as final.

Figure 3.2: Box plot of estimates across individual studies - first part



Note: This figure shows the first part of a box plot, where the reported estimates are grouped at the study level. The first 60 studies from the dataset are displayed in alphabetically ascending order. The red line represents the average effect across the literature. Each box's length represents the interquartile range between the 25th and 75th percentiles. The dividing line within each box indicates the median value. The whiskers extend to the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. Outliers are depicted as blue dots. The red line depicts the mean of the effect within the data. The data is winsorized at 1% level.

Figure 3.3: Box plot of estimates across individual studies - second part



Note: This figure shows the second part of a box plot, where the reported estimates are grouped at the study level. Fifty-five remaining studies from the dataset are displayed in alphabetically ascending order. The red line represents the average effect across the literature. Each box's length represents the interquartile range between the 25th and 75th percentiles. The dividing line within each box indicates the median value. The whiskers extend to the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. Outliers are depicted as blue dots. The red line depicts the mean of the effect within the data. The data is winsorized at 1% level.

Chapter 4

Publication bias

An investment in knowledge pays the best interest.

— Benjamin Franklin

It is widely acknowledged that attending school brings numerous benefits for one's future. To claim the opposite would simply appear foolish, considering the vast quantities of existing literature that explore the positive impact of schooling (Oreopoulos & Petronijevic 2013; Ritchie & Tucker-Drob 2018; Heckman et al. 2010; Psacharopoulos & Patrinos 2004). However, what happens if a researcher conducts an experiment, and the results suggest that education hurts the prospects of the subjects that took part? Such an experiment will most likely be viewed skeptically, if not frowned upon. The initial response of the publishers presented with such results might, in most cases, go more along the lines of "Perhaps there is something wrong with your setup?" rather than "Truly, what a revolutionary discovery!" In expectation of such a response and considering the time and often money invested into the experiment, the researcher is posed with a tough decision - keep the results as is, or sacrifice legitimacy in return for better publication prospects?

The issue described above is commonly referred to as publication bias and is exactly what this part of the paper explores. Among the many forms this malpractice can take, perhaps two are the most prominent. Firstly, studies can remain unpublished due to the discrepancy between their results and the existing knowledge, also known as the *file drawer problem* (Stanley 2005). Secondly, the results within those studies may be modified to gain higher order of statistical significance; this can be done by modifying the standard error or even the effect itself - a form of malpractice sometimes referred to as p-hacking (Simmons et al. 2011). Luckily, this manipulation can be detected within the data

using various statistical tests, given the unnatural patterns that the p-value distribution tends to exhibit in case such practices are employed.

Regarding publication bias in the literature on returns to education, five of the ten existing meta-analyses address the issue, as mentioned in Chapter 2. Out of these five studies, three detect the presence of publication bias, while two do not. These conflicting results leave a lot to be desired. For this, I find it crucial to shed more light on the publication bias issue and hope to bring further vital evidence.

In the rest of this chapter, I will first graphically explore the relationship between the effect and the standard effect using a simple visual test. Next, I will conduct multiple linear and non-linear tests to determine whether a quantifiable link exists between the two variables. I will also employ methods that do not assume any prior form of the relationship and test for structural breaks in the distribution of t-statistics in the data. Lastly, I will bring three completely new methods into the picture. These should help me detect p-hacking in the data and link the results together across multiple models with means of model averaging.

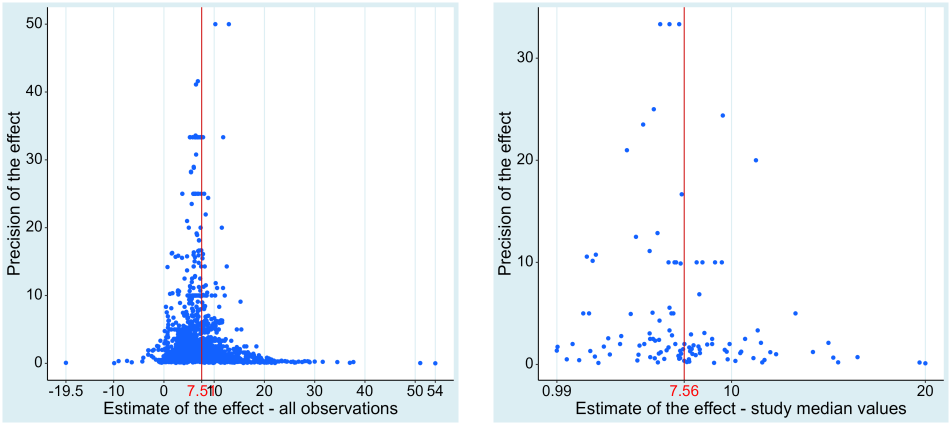
4.1 Funnel Plot

I first test for publication bias using the funnel plot (Egger et al. 1997; Stanley 2005). The genius of the method lies in its simplicity, where the main effect is plotted on the x-axis against a measure of precision on the y-axis. Usually (and in this case, too), the precision is calculated simply as the inverse of the standard error. Although Stanley (2005) suggests that alternatives can be used, such as the square root of the degrees of freedom, I opt for the standard error. After the plot is constructed, the most precise estimates should be clustered around the true effect mean, assuming that the data contains no publication bias, systematic heterogeneity, or small-sample effects. As precision decreases, the estimates get more scattered, creating an inverted funnel shape. In this shape, gaps or holes hint that data tampering exists within the literature.

As mentioned above, I construct the funnel plot using the inverted standard error as the measure for precision because all estimates in the dataset have their standard error reported (this was one of the conditions during data collection, as described in Chapter 3). Apart from a funnel plot with all collected data points, I also present a figure that displays only the medians of the effect for all 74 studies. These two graphs appear in the sub-figures of Figure 4.1.

Figure 4.1: The funnel plot shows no immediately suspicious patterns

(a) All observations (b) Study medians



Note: This figure displays two funnel plots as per Egger et al. (1997), where the percentage returns to an additional year of schooling are plotted on the x-axis against precision on the y-axis, measured as $1/SE$ (Standard Error). Plot (a) shows the funnel plot for all observations within the data (1754 data points), while plot (b) shows only the medians of each study (115 data points). The red line marks the mean of these data points. In case of no publication bias, these funnel plots should be symmetrically centered around the true mean.

No apparent asymmetry nor holes appear at first glance in the plot with all observations. The less telling plot with study medians exhibits a little more "emptiness" at places, but we can take this simply as a cause of the lower observation count. The crucial takeaway from the latter plot is the lack of suspicious outliers in either direction. Despite this relative consistency, both graphs are perhaps a little more dispersed for higher precision values than the ideal shape might have it. In any case, this is far from enough evidence to claim the presence of publication bias and, contrarily, a hint that the data may be relatively normal.

4.2 Funnel Asymmetry Tests

The funnel plot itself, albeit a quick and easy way of detecting obvious publication bias, is still a less precise method that relies on *eye-balling* and subjective interpretation, both hardly rigorous ways of conducting research. To establish the results quantitatively and more robustly, I turn first to the Funnel Asymmetry Tests - Precision Effect Testing (FAT-PET) (Egger et al. 1997; Stanley 2005; 2008).

These techniques test for the funnel plot asymmetry using a simple equation that regresses the effect on its standard error to uncover any correlation between the two. If such a correlation exists, it can be interpreted as a systematic

relationship between the effect and its standard error, indicating publication bias. Algebraically, the relationship can be written as:

$$S_{ij} = \beta_0 + \beta_1 * (SE_S)_{ij} + u_{ij}, \tag{4.1}$$

where S represents the returns to schooling effect for the i-th observation of the j-th study in the dataset, and SE_S corresponds to the effect's standard error. The slope coefficient, β_1, then measures the publication bias in the data, while the intercept coefficient, β_0, captures the "true" effect of returns to schooling corrected for publication bias. u_{ij} stands for the error regression term. In the tables below, I will refer to the slope coefficient with the label *Publication bias*, while the intercept will be labeled as *Effect beyond bias*.

If no publication bias is present in the data, the slope coefficient will be either 0 or close to it. Conversely, higher absolute values would indicate the opposite correlation between the effect and its standard error, thereby suggesting publication bias is present in the data. This is motivated by the assumption that both the effect and its standard error should be, statistically speaking, drawn from an independent, statistically symmetrical distribution. However, practically speaking, this is rarely the case.

The results of the funnel asymmetry tests can be viewed in Table 4.1. Firstly, I include a simple OLS model, followed by two models accounting for unobserved heterogeneity in the form of Fixed effect and Random effect estimators. Lastly, I introduce two models that weigh the equation, first by the inverse of the number of observations reported per study and second by precision. The motivation behind the last two models is to account for unobserved heterogeneity and heteroskedasticity, respectively. Four out of five of these methods find a statistically significant presence of publication bias, and all claim the underlying effect lies within the range of 6 and 7 percent, specifically between 6.294 and 6.708 percent. This indicates that the underlying effect might be slightly lower than the simple estimates' average, approximately by one percentage point. Furthermore, the lowest predicted value can be associated with the study-size weighted model (6.294), suggesting perhaps that studies of larger size drive the effect upwards. However, it is essential to note that this difference is relatively small compared to the other estimates. The discrepancy between the study-size weighted model and the fixed-effects model, the latter of which predicts the highest returns to education at 6.708 percent, is less than half a percentage point.

Table 4.1: Linear tests for publication bias

	OLS	**Study**	**Precision**
Publication Bias	0.832**	1.169***	0.262
(Standard Error)	(0.097)	(0.121)	(0.425)
Bootstrapped CI (PB)	[0.624, 1.035]	[0.92, 1.405]	[-0.833, 1.091]
Effect Beyond Bias	6.408***	6.294***	6.54***
(Constant)	(0.118)	(0.153)	(0.168)
Bootstrapped CI (EBB)	[6.164, 6.639]	[6.04, 6.645]	[6.189, 6.918]
Total observations	1,754	1,754	1,754
	FE	**BE**	**RE**
Publication Bias	0.746***	0.752***	0.747***
(Standard Error)	(0.06)	(0.244)	(0.058)
Bootstrapped CI (PB)			[0.514, 0.995]
Effect Beyond Bias	6.517***	6.741***	6.708***
(Constant)	(0.107)	(0.418)	(0.294)
Bootstrapped CI (EBB)			[6.398, 6.965]
Total observations	1,754	1,754	1,754

Note: The table displays the results obtained from estimating Equation 4.1 OLS = Ordinary Least Squares. FE = Fixed Effects. BE = Between Effects. RE = Random Effects. Precision = Estimates are weighted by the inverse standard error. Study = Estimates are weighted by the inverse number of observations reported per study. Standard errors, clustered at the study level, are included in parentheses. ***p<0.01, **p<0.05, *p<0.1

4.3 Non-linear Tests

The relationship between the effect and its standard error, as described in Equation 4.1, is assumed to be linear in the funnel asymmetry tests. However, it is crucial to acknowledge that this assumption does not always hold. In cases where the relationship behaves less straightforwardly, the FAT-PET tend to underestimate the underlying effect if it differs from zero (Stanley & Doucouliagos 2014). In the context of my data, this concern may be valid since most of the data points are positive and occasionally even reach double digits. To address potential non-linear forms of the relationship that may appear in the data, I present six techniques that relax the linearity assumption.

The first is the Weighted Average of Adequately Powered (WAAP), introduced by Ioannidis et al. (2017). Their proposition involves the application of unrestricted Weighted Least Squares (WLS) only on observations of adequately powered studies. "Power" here refers to a study's ability to detect whether an effect if it is truly present in the data. The more power a study has, the bigger its reliability. In technical terms, the power is calculated using the statistical significance of estimates and then compared to their standard errors. As per

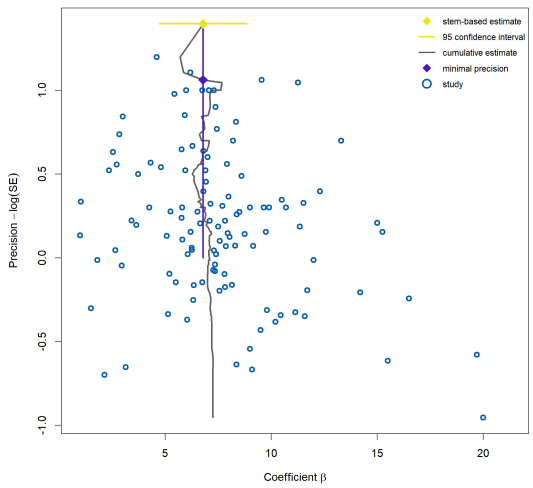


Figure 4.2: Stem-based method

Note: This figure displays a non-linear estimation of the true effect (Furukawa 2019). The main effect, labeled *Coefficient* β, is plotted on the x-axis against precision, calculated as $log(SE)$ (Standard Error). The yellow diamond and the whiskers of the same color denote the 95% confidence interval of the stem-based estimate of the effect. The dark grey line indicates the predicted estimates throughout various levels of data, while the purple diamond shows the minimal precision value above which lies the stem. The blue circles then represent the individual effect estimates.

the original paper, I kept only the estimates of studies that display power over 80%. Strikingly, 1469 out of the 1754 estimates in my dataset get identified as adequately powered. Using these estimates, WAAP then proposes an estimate of 6.9% which is slightly higher than any of the linear models presented in Table 4.1.

The second approach, proposed by Stanley et al. (2010), entails discarding 90% of data and keeping only the top 10 percent with the highest precision. This somewhat paradoxical approach stems from the idea that most researchers use statistical significance as the primary benchmark for deciding whether to publish the estimate. Stanley et al. (2010) show that if most of the less precise estimates are discarded, the publication bias within the data sample drops considerably. In my data, 10% of estimates correspond to 176 observations. The Top10 model yields a modest result of 6.439%.

Furukawa (2019) chooses a similar tactic by selecting a specific number of the most precise estimates. The cutoff is determined by minimizing mean square error; this aims at striking a balance between variance/efficiency and bias. The selected points from what is referred to as "stem" are those with the lowest mean square error. You may find a visual representation of this method in Figure 4.2. The coefficient of returns to education suggested by this approach is 7.2%, the highest among all the proposed regression-based results thus far.

Further, I estimate the Hierarchical Bayes model by Allenby & Rossi (2006). The procedure employs Bayesian statistics to leverage variability within individual studies to determine the weights of individual observations aggregated at the study level. The model parameters get treated as random variables in-

stead of fixed numbers, allowing for variability at multiple levels within the dataset. As such, different units can have comparable sharing strength, allowing for more robust estimates. The hierarchical part stems from the fact that priors are specified using another model (called hyperprior) instead of a direct specification, as is usually done in Bayesian modeling. This complex multi-level modeling framework yields an estimate of 6.8% in my case. Additionally, the analysis suggests the presence of publication bias at a significance level of 1%.

The next test is the Selection model proposed by Andrews & Kasy (2019). The authors argue that the publication probability for estimates remains constant at similar levels of statistical significance, a concept called "conditional publication probability." Once a certain statistical significance threshold is crossed, the publication probability changes. Andrews & Kasy (2019) then demonstrate how this probability can be calculated non-parametrically, and utilizing the inverse of this probability as new weights, they obtain a non-biased distribution of the estimates. Using a t-distribution at the 5% significance level (cutoffs for $p(.)$ set to 1.96), I obtain the result of 6.548%. The method also proposes that estimates at the 5% significance level are more likely to be published than insignificant ones (P = 2.764).

As the last of the non-linear techniques testing for publication bias, I add the Endogenous Kink (EK) model introduced by Bom & Rachinger (2019). Using the argument that the publication bias is usually absent for sufficiently large studies, the EK model finds a cutoff value below which publication bias should not appear. Bom & Rachinger (2019) then fit a piecewise linear regression with

Table 4.2: Nonlinear tests for publication bias

	WAAP	**Top10**	**Stem**	**Hier**	**AK**	**Kink**
Publication Bias				0.504***	2.764***	0.262
(PB SE)				(0.165)	(0.112)	(0.39)
Effect Beyond Bias	6.9***	6.439***	6.783***	6.801***	6.548***	6.54***
(EBB SE)	(0.092)	(0.146)	(1.055)	(0.269)	(0.091)	(0.054)
Total Observations	1,754	1,754	115	1,754	1,754	1,754
Model observations	1,469	176				

Note: The table reports estimates of the effect beyond bias using six non-linear methods and estimates of the publication bias obtained using two of these methods. WAAP = Weighted Average of the Adequately Powered (Ioannidis et al. 2017). Top10 = Top10 method by Stanley et al. (2010). Stem = the stem-based method by Furukawa (2019) where P represents the probability of results insignificant at 5% being published relative to the probability of the significant ones at the same level. Hier = Hierarchical Bayes model (Allenby & Rossi 2006). AK = Andrews & Kasy (2019)'s Selection model. Kink = Endogenous kink model by Bom & Rachinger (2019). Standard errors, clustered at the study level, are included in parentheses. ***p<0.01, **p<0.05, *p<0.1

a kink at this cutoff point, allowing non-linearity in the model. An advantage of this approach is that this method reduces to a simple linear model as the effect approaches zero, where said linear methods perform well. As such, the EK approach should provide more robust results than its linear counterpart. In my case, the suggested value of the main effect is 6.54%, which falls right into the average of the rest of the (both linear and non-linear) results. The model also provides a non-significant estimate of the presence of publication bias. This marks the last of non-linear methods; all of the results obtained from these estimations can be found in Table 4.2.

All but one of the six models propose a statistically significant effect beyond bias within the 6 to 7 percent range. Only the stem-based method suggests a coefficient for returns to education above 7% (specifically, 7.2%). These results align with the linear approach and confirm the behavior observed thus far. The Hierarchical Bayes indicates a strong presence of publication bias, while the Endogenous Kink method result is insignificant. Finally, the Selection model proposes that results at the 5% significance level have a considerably higher chance of publication than insignificant ones.

4.4 Tests Without the Exogeneity Assumption

Until now, the publication bias tests have been based on the assumption that the correlation between the effect and the standard error indicates publication bias. However, this introduces, by definition, endogeneity into the equation. To see how this issue can be treated, it is essential to understand how it arises in the first place.

The correlation in the data, and thus endogeneity, can come from several sources. First, it could be a simple measurement error or wrong calculation procedure that introduces correlation into the data; the standard error, too, is an estimate, after all. Second - this is what the publication bias gets associated with perhaps the most - the endogeneity may arise from a conscious and deliberate tampering of the standard error to improve significance. And lastly, any unobserved heterogeneity may also introduce correlation, this time in the form of inherent methodological differences that may systematically influence the results. To display the estimate-error relationship clean of endogeneity, I utilize two techniques - IV regression and p-uniform* (van Aert & van Assen 2021).

First, the IV regression, for which we need an instrument. The criteria for

finding a valid one are relatively simple - it should be a metric that somehow captures the behavior of (correlates with) standard error while having no relationship to the estimate. Using such metric, it should be possible to derive the publication bias coefficient (β_1 from Equation 4.1) not poisoned by endogeneity. Several instruments appear valid here, including $\frac{1}{\sqrt{n_{\text{obs}}}}$, $\frac{1}{n_{\text{obs}}}$, $\frac{1}{n_{\text{obs}}^2}$, and $\log(n_{\text{obs}})$, where n_{obs} stands for the number of observations associated with each estimate. The number of observations variable holds several inherent properties that make all these instruments valid options. Firstly, the size of an experiment, or the number of subjects in the study, does not directly change the population-wide effect. If such a true effect exists, it should be independent of how many subjects we include in the analysis. Secondly, the standard error decreases as the sample size increases. This is a fundamental principle of statistics. In other words, the more subjects there are in the study, the bigger the confidence that the findings based on that sample are close to the results had the whole population been used for calculation.

Still, which of these four proposed instruments is the best? To find out, I wrote a helper function in R that automatically detects the best-performing instruments based on the results of several specification tests. These are, namely, the Underidentification test, the Weak identification test, the Stock-Yogo weak ID test, and the Sargan statistic.[1] I omit the numeric results of these tests as they are not crucial for interpreting the results, and only mention that $\frac{1}{\sqrt{n_{\text{obs}}}}$ performed the best out of the four instruments. Using this instrument, the IV regression gives 6.155% as an estimate of returns to education, which coincides with the estimates computed up to this point.

As another way of estimating the effect-error relationship without prior assumptions about its form, I turn to the p-uniform* method. This approach, proposed by van Aert & van Assen (2021), builds on the p-uniform method (Van Aert et al. 2016). The core idea stems from the principle that the p-values in the data should be uniformly distributed at the true effect size. This line of thinking requires no assumptions about the form nor correlation of the relationship and helps search for publication bias in a novel way. The p-uniform* method, then, improves the p-uniform approach in efficiency, precision, and between-study variance detection. In my data, this technique estimates the effect to be 9.52% and indicates the presence of publication bias, both at high levels of significance. Results of both methods can be found in table Table 4.3.

[1] All of these specification tests are in-built into the *ivreg* function of the ivreg R package, which I used to estimate this method. Source here.

Table 4.3: Relaxing the exogeneity assumption

	IV	**p-uniform***
Publication Bias	1.295***	L = 9.439
(PB SE)	(0.281)	(p = 0.002)
Effect Beyond Bias	5.813***	9.52***
(EBB SE)	(0.354)	(3.291)
Observations	1,754	1,754
F-test	29.153	

Note: IV = Instrumental Variable Regression; one over the square root of the number of studies is used as an instrument for the standard error. Standard errors, reported in parentheses, are also clustered at the study level. p-uniform* = method proposed by van Aert & van Assen (2021); L represents the publication bias test t-statistic; the corresponding p-value can be found in parentheses. ***p<0.01, **p<0.05, *p<0.1

While the instrumental variable approach proposes rather sensible results, the p-uniform* is an outlier among previous estimates. This is perhaps more perplexing given that, were between-study to cause the effect's overinflation, p-uniform* is a method that should account for this. Among various possibilities, these results may stem from a calculation error or, perchance, a hidden trend or anomaly within the data, which is hard to detect. Suffice it to say I dug into the calculation multiple times to validate that all specifications and other inputs were sensible; still, I could not find anything out of the ordinary. As such, I present the results with a grain of salt but believe them to be fully valid.

4.5 Caliper Tests

Yet another method I will use to search for deviations from normality in the reported literature results are Caliper tests, developed by Gerber et al. (2008). Their proposed approach does not assume any prior relationship between the effect and the standard effect, similar to the tests from Section 4.4. Here, t-statistics are subjected to scrutiny, and the authors argue that upon looking at the immediate vicinity of a conventional significance level, no structural breaks in the distribution of t-statistic should occur. In other words, the t-statistic distribution should, in theory, behave relatively normally, and any large jumps may indicate the presence of publication bias.

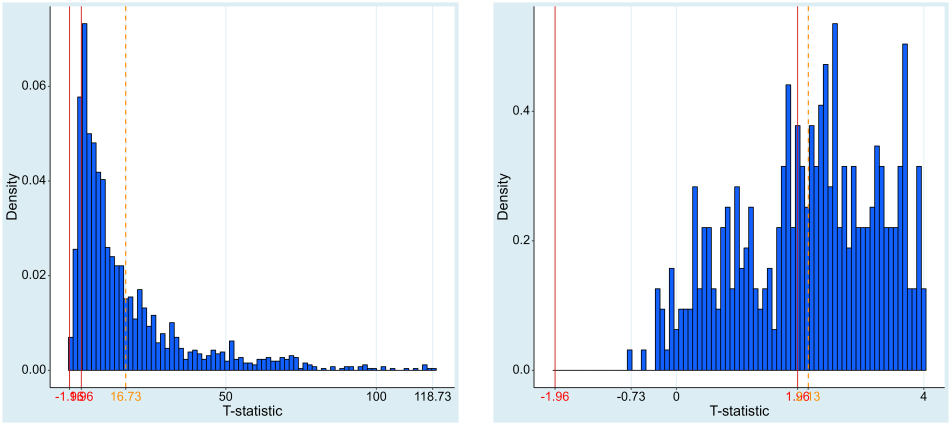
Going into more detail, Gerber et al. (2008) suggest observing the number of t-statistics around significant t-statistic thresholds, such as 1.69 or 1.96, in intervals of varying widths, called Caliper widths. If, within any half of that

interval, there is a significant imbalance in the number of t-statistics compared to the other half, it indicates a structural break around the observed threshold. In my case, I will explore how the t-statistics included from all studies of the data set behave around thresholds 1.645, 1.96, and 2.58, with Caliper widths of 0.05, 0.1, and 0.2. The choice of the latter is arbitrary, while the choice of the former stems from the fact that the three values correspond to the 1%, 5%, and 10% significance levels. In academia, it is a common practice to append asterisks to results when presenting estimates together with their standard errors and hence, t-statistics. Unfortunately, this practice inadvertently emphasizes results marked with these asterisks (Simmons et al. 2011). As such, researchers may be tempted to include these asterisks in their tables at the cost of honesty, leading them to tamper with their figures (most notably standard errors). Consequently, publication bias may arise.

In Figure 4.3, you may find the distribution of t-statistics in my data, while Table 4.4 reports the results of Caliper tests described in the previous paragraphs.

Two quick notes about the results are in order. First, there are very few (only 34 out of the 1754 observations) estimates with negative t-statistics. Looking at the distribution from a purely statistical standpoint, it appears peculiar that the other 1730 are all associated with a positive t-statistic. From

Figure 4.3: The distribution of t-statistics is heavily skewed



Note: The figure depicts the distribution of t-statistics associated with estimates within the dataset. Plot (a) shows all t-statistics in the dataset, while plot (b) focuses on a close up view around 0. The two red lines mark the critical significance values -1.96 and 1.96 (from left to right) at the 95% confidence level. The dotted orange line represents the mean t-statistic within the data. Outliers are hidden for clarity of presentation, but we included them in the calculations.

Table 4.4: Caliper tests at values 1.645, 1.96 and 2.58

	Threshold 1.645	**Threshold 1.96**	**Threshold 2.58**
Caliper width 0.05	0.517***	0.243***	0.152***
SE	(0.084)	(0.063)	(0.046)
Observations	7	17	18
Caliper width 0.1	0.467***	0.23***	0.183***
SE	(0.069)	(0.051)	(0.037)
Observations	13	25	28
Caliper width 0.15	0.483***	0.269***	0.186***
SE	(0.042)	(0.041)	(0.028)
Observations	26	37	45

Note: The table shows the results of three sets of Caliper tests by Gerber et al. (2008) These sets are carried out around t-statistic thresholds of 1.645, 1.96 and 2.58, which correspond to the 1%, 5%, and 10% t-statistic significance levels. Caliper width denotes the width of the interval around the t-statistic, e.g., Caliper width 0.05 for threshold 1.96 means $t \in < 1.91; 2.01 >$. A test statistic of 0.243 means that roughly 74% of estimates appear above the threshold and roughly 26% below it. SE = Standard Error, Observations = Total number of observations in the interval around the threshold. Standard errors, clustered at the study level, are included in parentheses.

a practical perspective, however, this makes a lot of sense if we presume that the true effect indeed lies around 7%. This assumption appears quite feasible, given the consistency of the tests carried out in the previous sections.

The second note should be addressed to the results of the Caliper tests. The jumps around thresholds could be described as *striking*, *considerable*, and *mild*, talking about the 1%, 5%, and 10% thresholds, respectively. Speaking more bluntly, the words *high*, *medium*, and *low* could be used. The t-statistics just above the thresholds of 1.645 and 1.96 are being over-reported in the data sample to some degree. So far, we have obtained somewhat skeptical views on publication bias presence in the dataset, but perhaps these thresholds could represent initial tangible indications of reporting misbehavior.

4.6 Novel Tests for Detecting Publication Bias

As the last chapter of my hunt for publication bias, I present three new methods that further explore the issue of publication bias. The first two methods deal with p-hacking and have been developed very recently. They are, in order, the Elliott tests by Elliott et al. (2022) and the Meta-Analysis Instrumental Variable Estimator (MAIVE) estimator by Irsova et al. (2023). While the former paper analyzes the distribution of p-values across different studies, the latter focuses on the issue of spurious regression and how p-hacking precision can produce biased results. As the last new method, I add Robust Bayesian Model

Averaging (RoBMA) (Maier et al. 2022), a technique that can produce results of unparalleled quality and precision (Bartoš et al. 2023).

First, let us talk about the Elliott tests. Elliott et al. (2022) propose an approach where *no p-hacking in the literature* is considered as a null, and using a set of general assumptions, they test this hypothesis against an alternative of *p-hacking in the literature*. The p-curves for various subsets of the true effects should be non-increasing and continuous, providing p-hacking is absent. For p-values based on t-tests, the authors then devise a new set of assumptions under which the lack of p-hacking should lead to a monotonous form of the p-curve. The advantage of the method lies in the fact that no threshold for the t-statistic needs to be specified; the technique only focuses on the p-curves. In my case, I present the results of the two tests I vaguely described - the test for non-increasingness of the p-curve and the test for monotonicity and bounds. With sufficiently low p-values, we could reject the null hypothesis of no p-hacking, but that is not the case in my dataset. Both tests yield p-values over 0.8, so there is insufficient evidence to reject the null in favor of the alternative (p-hacking).

Next, the MAIVE estimator developed by Irsova et al. (2023). The authors argue that precision, one of the critical metrics in meta-analytic research, is prone to p-hacking. In their paper, Irsova et al. (2023) raise several points of concern regarding the metric. First, the author must calculate the metric

Table 4.5: P-hacking tests

Panel A: P-hacking tests by Elliott et al. (2022)	**Test for non-increasingness**	**Test for monotonicity and bounds**
p-value	0.819	0.871
Observations ($p \leq 0.1$)	1,610	1,610
Total observations	1,754	1,754
Panel B: MAIVE estimator (Irsova et al. 2023)	**MAIVE coefficient**	
Coefficient	5.736***	
Standard Error	(0.460)	
Observations	1,754	
F-test	12.491	

Note: This table shows the results of two techniques that detect p-hacking. Panel A shows the results of p-hacking tests by Elliott et al. (2022), namely the histogram-based test for non-increasingness and the histogram-based test for monotonicity and bounds. Panel B reports the results of the spurious precision robust approach using the MAIVE estimator by Irsova et al. (2023). F-test = Test statistic of the IV first step F-test. Cluster-robust standard errors are used in the MAIVE estimation. These are reported in parentheses. ***p<0.01, **p<0.05, *p<0.1

using reported standard errors; this makes the calculation easily 'p-hackable.' Second, even small amounts of p-hacking can profoundly impact the results. Precision is often used as a weighting metric in methods such as linear tests, plus it holds a vital role as one of the main axes of the funnel plot. As a remedy for this, Irsova et al. (2023) propose a new estimator utilizing the instrumental variable approach (MAIVE), where the reported variance is instrumented using the inverse sample size. This approach should help mitigate the impact of spurious precision in the data. This estimator suggests 5.736% percent returns to education, a figure lower than any of the tests conducted thus far. The F-statistic of 12.491 then shows the inverse sample size to be a good instrument for reported variance. The results of both p-hacking tests are shown in Table 4.5.

The last of the procedures exploring publication bias is the RoBMA by Maier et al. (2022). The idea lies in estimating multiple meta-analytic models and combining them using Bayesian model averaging. Each model is assigned a different weight, and individual components, such as the presence or absence of an effect, are tested using Bayes factors. In Table 4.6, I present two panels: the first panel displays the model-averaged estimates of the effect, while the second panel summarizes the individual components - effect, heterogeneity, and publication bias. The effect estimates propose a mildly confident claim that the effect lies just above 7% percent, which is slightly more positive than the estimates of both linear and non-linear models. Among the four models used

Table 4.6: Robust Bayesian Model Averaging

Panel A: Model-averaged estimates				
	Mean	**Median**	**0.025**	**0.975**
Coefficient	7.123**	7.122**	6.946**	7.299**
Standard Error	(3.506)	(3.504)	(3.373)	(3.645)
Observations	1,754	1,754	1,754	1,754
Panel B: Summary of individual components				
	Models	**Prior Prob.**	**Post. Prob.**	**Inclusion BF**
Effect	2/4	0.500	1.000	∞
Heterogeneity	2/4	0.500	1.000	∞
Bias	0/4	0.500	0.000	0.000

Note: This table shows the Robust Bayesian Model Averaging method by Maier et al. (2022). Panel A contains four descriptive statics of the estimates obtained from model-averaging - mean, median, 2.5th quantile, and 97.5th quantile. Standard errors are reported in parentheses. In Panel B, the summary of three individual components is displayed - effect, heterogeneity, and publication bias. Models = Probability of each model assuming a given individual component. Prior Prob. = Prior Probability. Post. Prob. = Posterior Probability. Inclusion BF = Inclusion Bayes Factor. ***p<0.01, **p<0.05, *p<0.1

to estimate individual components[2], the probability of a model assuming the presence of effect or heterogeneity is 2/4 (50%), while for publication bias it is 0/4 (0%).

To summarize, all models agree that schooling positively affects log wage (returns to an additional year of schooling of 5.736-9.520%), and most suggest it lies somewhere between 6 and 7 percent. The vast majority of results associated with these techniques are also highly statistically significant. As for publication bias, the story is a bit more tangled. Some linear and non-linear methods argue for its presence, while others are against it. Even when relaxing the assumption of exogeneity of the standard error, the results appear mixed. Novel methods almost uniformly suggest the lack of publication bias, apart from MAIVE, which predicts lower returns to education when instrumenting for study variance. Lastly, the Caliper tests show that sizeable jumps exist in the distribution of t-statistics around 1% and 5% significance levels. Perhaps too many cooks spoil the broth, so a single interpretation appears unfeasible, and I would suggest considering the results of the presented methods individually.

[2]I used the base specification of the *RoBMA* method. For the list of models and other parameters used, see the source code of the method, available here.

Chapter 5

Heterogeneity

Thus far, my analysis has focused primarily on the relationship between the true effect and its standard error. Several methods from the previous chapter, such as the IV regression, p-uniform*, or RoBMA, provided us with a quick glimpse into the topic of systematic heterogeneity. However, none delivered a more complex overview of the data's nature. This chapter aims to do precisely that - delve deeper into the study design and search for systematic patterns that may reveal more about the behavior of the effect. For this purpose, I will utilize two methods, Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA). These should help me identify the influence of different variables on the effect behavior and quantitatively capture the magnitude of this influence. Before constructing any models, however, it is crucial to explain and explore the dataset structure first.

5.1 Variables

I constructed the dataset aiming to comprehensively capture the most important categories that define the context of the collected data and the studies they come from. As such, I identified six categories, which I named as follows: the actual estimates along with their descriptive statistics, estimate characteristics, data characteristics, spatial/structural variation, estimation method, and publication characteristics. Across these six categories, I collected 37 distinct variable groups. Note that a group here could mean either a standalone variable (i.e., data year) or a group of variables (i.e., low/middle/high-income country). In the latter case, the variable groups consist either of dummies, or ratios, such as the ratio of subjects living in an urban area. The list of all quan-

tifiable, relevant variables can be found in Table 5.1. To keep visual clarity, I excluded variables that could not be easily quantified, such as the country where the study was conducted, or variables irrelevant to the effect behavior explanation, such as observation id.

Table 5.1: Definition and summary statistics of regression variables

Variable	Description	Mean	SD
Effect	The effect of an additional year of schooling on logarithmic wage.	7.476	4.439
Standard Error	The standard error of the main effect.	1.284	1.693
Estimate characteristics			
Estimate: City	=1 if the estimates within the study can be aggregated on a city level.	0.119	0.323
Estimate: Sub-region	=1 if the estimates within the study can be aggregated on a subregional level.	0.099	0.299
Estimate: Region	=1 if the estimates within the study can be aggregated on a regional level.	0.309	0.462
Estimate: Country	=1 if the estimates within the study can be aggregated on a country level.	0.395	0.489
Estimate: Continent	=1 if the estimates within the study can not be aggregated on a country level or smaller (reference category).	0.079	0.269
Data characteristics			
Study Size	The logarithm of the number of estimates collected from the study.	2.942	0.637
Yrs. of Schooling	The average number of years of schooling attained by the subjects.	11.116	3.461
Yrs. of Experience	The average number of years of experience attained by the subjects.	18.351	7.450
Education: Years	=1 if authors report schooling in years.	0.634	0.482
Education: Levels	=1 if the authors report schooling in levels (e.g., attained college degree) (reference category).	0.366	0.482
Wage: Log Hourly	=1 if the dependent variable in the regression is log hourly wage.	0.531	0.499
Wage: Log Daily	=1 if the dependent variable in the regression is log daily or weekly wage.	0.095	0.293
Wage: Log Monthly	=1 if the dependent variable in the regression is log monthly wage.	0.211	0.408
Wage: Annual Earnings	=1 if the dependent variable in the regression is log of mean annual earnings (reference category).	0.162	0.369
Micro Data	=1 if the study uses micro data.	0.177	0.382
Survey Data	=1 if the study uses data from a survey.	0.534	0.499
National Register Data	=1 if the study uses data from a national register (reference category).	0.289	0.453
Cross-sectional Data	=1 if the study uses cross-sectional data.	0.361	0.481
Panel Data	=1 if the study uses panel data (reference category).	0.639	0.481
Data Year	The logarithm of the average year of the study's time span	7.599	0.006
Spatial/structural variation			
No Education	The percentage of subjects that attained no education (reference category).	0.126	0.148
Primary Education	The percentage of subjects that attained only primary education.	0.177	0.151
Secondary Education	The percentage of subjects that attained only secondary education.	0.388	0.196
Higher Education	The percentage of subjects that attained any form of higher education.	0.309	0.247
Wage Earners	The ratio of wage earners to self-employed subjects in the study (= 1 if wage earner, = 0 if self-employed).	0.837	0.205
Self-Employed	The ratio of self-employed to wage earners subjects in the study (= 1 if self-employed, = 0 if wage earner) (reference category).	0.163	0.205

Continued on next page

Table 5.1: Definition and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Male	The ratio of male to female subjects in the study (= 1 if male, = 0 if female).	0.650	0.350
Female	The ratio of female to male subjects in the study (= 1 if female, = 0 if male) (reference category).	0.350	0.350
Private Sector	The ratio of private to public sector workers (= 1 if private sector worker, = 0 if public).	0.596	0.163
Public Sector	The ratio of public to private sector workers (= 1 if public sector worker, = 0 if private) (reference category).	0.404	0.163
Ethnicity: Caucasian	The ratio of Caucasian to non-Caucasian subjects in the study (= 1 if Caucasian, = 0 if not).	0.227	0.419
Ethnicity: Other	The ratio of non-Caucasian to Caucasian subjects in the study (= 1 if non-Caucasian, = 0 if Caucasian) (reference category).	0.773	0.419
Rural	The ratio of rural to urban workers (= 1 if rural worker, = 0 if urban).	0.297	0.191
Urban	The ratio of urban to rural workers (= 1 if urban worker, = 0 if rural) (reference category).	0.703	0.191
Reg: Advanced Econ.	=1 if the study was conducted in a country with advanced economy. (reference group)	0.498	0.500
High Income Countries	=1 if the study was conducted in a high income country (reference category)	0.507	0.500
Median Expenditure	The logarithm of the median expenditure in the country in a given year.	8.584	1.420
Minimum Wage	The logarithm of the minimum wage in the country in a given year.	5.853	1.536
Academic Freedom Index	The academic freedom index reported for the country in a given year.	0.712	0.266
Estimation method			
Method: OLS	=1 if the authors use Ordinary least squares (reference category).	0.664	0.473
Method: Cohort/FE	=1 if the authors use Cohort-type or Fixed-effects estimation.	0.058	0.234
Method: 2SLS	=1 if the authors use Two-Stage least squares estimation.	0.095	0.294
Method: Heckman	=1 if the authors use Two-step estimation (Heckman and Polachek, 1974).	0.062	0.240
Method: Probit	=1 if the authors use Probit estimation.	0.022	0.147
Method: IV	=1 if the authors use Instrumental variables estimation.	0.111	0.314
Ability: Direct	=1 if the authors include a direct measure of ability in their study.	0.135	0.341
Ability: Proxied	=1 if the authors use a proxy for ability in their study.	0.204	0.403
Ability: Uncontrolled	=1 if the authors acknowledge, but do not control for ability in any way in their study.	0.425	0.494
Ability: Unmentioned	=1 if the authors do not mention ability anywhere in their study (reference category).	0.223	0.417
Control: Age	=1 if the authors control for age in the regression.	0.344	0.475
Control: Age^2	=1 if the authors control for age in quadratic form in the regression.	0.275	0.447
Control: Experience	=1 if the authors control for experience in the regression.	0.607	0.489
Control: Experience^2	=1 if the authors control for experience in quadratic form in the regression.	0.512	0.500
Control: Ethnicity	=1 if the authors control for ethnicity in the regression.	0.251	0.434
Control: Health	=1 if the authors control for health in the regression.	0.135	0.342
Control: Gender	=1 if the authors control for gender in the regression.	0.367	0.482
Control: Marriage	=1 if the authors control for marriage in the regression.	0.361	0.480
Control: Occupation	=1 if the authors control for occupation of the subjects in the regression.	0.142	0.349
Control: Firm Char.	=1 if the authors control for firm characteristics in the regression.	0.149	0.357
Control: Area	=1 if the authors control for area type in the regression (e.g., urban, rural).	0.418	0.493
Control: Macro Var.	=1 if the authors control for macroeconomic variables in the regression.	0.347	0.476

Continued on next page

Table 5.1: Definition and summary statistics of regression variables (continued)

Variable	Description	Mean	SD
Publication characteristics			
Impact Factor	The logarithm of the Journal Citations Report impact factor of the study (as of January 2023; = 0 in case of no publication).	-0.906	1.533
Citations	The logarithm of the mean number of Google Scholar citations received per year since the appearance of the study in Google Scholar (as of January 2023).	4.029	2.177
Study: Published	=1 if the study was published in a journal.	0.764	0.425
Study: Unpublished	=1 if the study was not published in a journal (reference category).	0.236	0.425
Publication Year	The logarithm of the number of years between the publication (or issuing) of this study and the publication year of the earliest published study in the sample.	3.332	0.339

Note: This table presents the summary statistics and descriptions for various study characteristics eligible for inclusion in Bayesian Model Averaging. Variables marked as *reference categories* were automatically excluded from the procedure, as this would create a dummy variable trap. SD = standard deviation, OLS = Ordinary Least Squares, FE = Fixed Effects, 2SLS = 2 Stage Least Squares, IV = Instrumental Variable.

Let us take a closer look at five of the six[1] variable categories and try to understand the reasoning behind my choices of this particular variable setup.

5.1.1 Estimate Characteristics

There are only a handful of variables that I identified as vital as far as effect characteristics are concerned. Moreover, variables such as the number of observations, or degrees of freedom, are not telling enough to be included in the model averaging. As such, the only full-fledged variable group included in this category is the estimate type, when divided into the size of the region. The estimates of over 70% of studies in the dataset can be clustered into regional or country levels. Examples of such studies include Walker & Zhu (2008); Fang et al. (2012), or Angrist & Krueger (1991). Sporadically (Krafft et al. 2019; Chanis et al. 2021), the authors focus on the city/sub-region level estimates or aggregate their results at a level of a continent or a group of countries.

5.1.2 Data Characteristics

Two variables are perhaps the most important in the category of data characteristics - *years of schooling* and *years of experience*. These represent the founding blocks of the Mincer equation and can be linked together using the age of subjects as described in Equation 2.2. Across all studies in the data, the average reported number of schooling years equals 11.116, while 18.351 represents the

[1]The statistical properties of the estimate have already been described in Chapter 3.

average reported experience of subjects. Given that 781 observations in the data (roughly 44% of all observations) are not directly reported, the *years of experience* variable may be inflated by the calculation. Indeed, upon removing all observations that had to be manually calculated using Equation 2.2, the average years of experience in the sample drops to 15.637. Nonetheless, to the best of my knowledge, there is no other way to circumvent this shortcoming. Consequently, I will use the reported number of 18.351 in further calculations. To see examples of studies that fail to report years of schooling and/or experience, see Pischke & von Wachter (2005); Psacharopoulos (1982), while for studies that report both, see Belzil & Hansen (2002); Girma & Kedir (2005).

Another crucial variable captures how education is reported - years or levels[2] In about two-thirds of all studies, years of attained education is used instead of the highest attained level (primary school, secondary school, etc.). It should be noted here that in cases a study reported both types, but the results captured the same outcome, I chose to collect only the number of years and discard the estimates in levels. This is to avoid collecting duplicate results. Harmon et al. (2002) is an excellent example of a study that utilizes reporting of schooling years, while Duraisamy (2002) provides a counterexample.

The last variable worth a mention from this category is the variable denoting cross-section/panel data. Initially, I coded a short/long run variable under the *estimate characteristics* that divided studies according to their run-time into those of length above and below one year. However, after the collection, I found that the cross-section/panel variable almost entirely captured this information, so I kept only this variable in the data. Nearly two-thirds of the collected experiments work with panel data such as longitudinal surveys (see Harmon et al. (2003)). On the other hand, one-third of them deal with cross-sectional data (Lemieux & Card (2001) as an example).

The rest of the variables in this category is self-explanatory. For the complete list, see Table 5.1.

5.1.3 Spatial/Structural Variation

A whole array of variables that capture study variation are all coded under the category *spatial/structural variation*. In most cases, this refers to either characteristics of the study subjects or the country in which the study is conducted. Pointing out a handful of crucial statistics that tie to these variables, we can

[2]See Section 3.2 for more details about this classification.

see that most of the data sample consists of wage workers (83.7%), 65% of the subjects are male, 22.7% come from the Caucasian ethnicity, 70.3% live in the urban area, about half of them (49.8%) come from a country with an advanced economy, and their average age is 35.69. To see the rest of the statistics, see Table 5.1.

Most of the choices regarding the variables themselves should be more or less straightforward. As such, I would like to focus on the calculation behind some of these instead. For example, the variables *median expenditure* and *minimum wage* are notably coded on the country-year level, meaning a data point exists for every unique country-year pair. This is to account for country-level heterogeneity, as well as inflation. Another variable, the *academic freedom index*, is too coded in this way.

Some variables, such as the rural/urban sector, are set up as ratios. Paweenawat & Vechbanyongratana (2015), for example, report exactly 47.4% of subjects that live in rural areas, and 57.6% that live in urban ones. This variable structure allows us to retain more information while behaving as a simple dummy in case only one of the alternatives is present in the data, such as when all subjects live in a city. I also employ this ratio-type setup with multiple categories in the variable that denotes the highest attained education. Here, the choices are split between primary, secondary, and higher education, as well as no education. When the authors report only several of these but not all, such as in the case of Chanis et al. (2021), I set the remaining variable categories to 0.

A more complex issue arises when more data points are missing, however. As an example, 32.5% of the 1754 studies do not report whether their subject pool consists of wage workers or self-employed individuals. 53.4% then omit the information on area type (urban/rural), and 60.5% fail to specify whether the subjects work in a private or a public sector. To run the model averaging, the dataset has to contain no missing points in the employed variables. As such, I resort to interpolation, whose specifics I explained earlier in Chapter 3.

5.1.4 Estimation Method

Regarding the actual estimation of the Mincer equation, the practices literature can be explained by three major variable sub-categories. Firstly, the estimation method used by the studies. Two-thirds of studies in the dataset (66.4%) use simple OLS for the estimation, while the rest use one of several other methods,

including the Fixed-effects, Probit model, Instrumental variable regression, or Two-stage least squares. Several studies, such as De Brauw & Rozelle (2008), employ a two-step estimation described in Heckman & Polachek (1974).

Secondly, the *ability* variable, described in Chapter 2, is also coded here. We can see that 13.5% of studies include ability directly in the regression, 20.4% use a proxy of some kind, 42.5% do not control for ability but are aware of it, and 22.3% do not mention ability or ability bias in any way.

Lastly, I add information on whether a study controls for variables such as age, experience, ethnicity, health, gender, marital status, etc. Usually, such as in the case of Girma & Kedir (2005) or Harmon et al. (2002), only one of the two variables of age and experience are included. On the other hand, the included variable comes very frequently with the squared term, as described in the original Mincer equation. As for the other controls, there seems to be no obvious pattern in the studies, and the authors appear to be choosing the controls arbitrarily based on their study goals, data availability, or personal preferences.

5.1.5 Publication Characteristics

The last of the variable categories that I chose to employ denotes various publication characteristics of the included studies. The number of Google Scholar citations, the year of publication, or the Journal Citations Report impact factor are among the handful of variables within this category. As described in Chapter 3, I collected all journal/study data at a single time point, namely in January 2023. Although it is possible that the status of several of the included studies changed from then, I still value direct comparability more than keeping the information up-to-date with the latest changes.

Interestingly, but perhaps not surprisingly, 76.4% of studies within the sample were published in a journal, and the mean number of citations for a study comes up to 56.2. This relatively high figure ties directly to the fact that roughly a third of the dataset consists of studies identified by snowballing - an activity aimed at targeting the most relevant and well-established relevant papers on the topic. Understandably, all of these papers have attained publication status, or their credibility would not be established.

With the variable setup out of the way, we can move on and employ these variables in exploring the effect behavior in a more detailed way.

5.2 Model Averaging

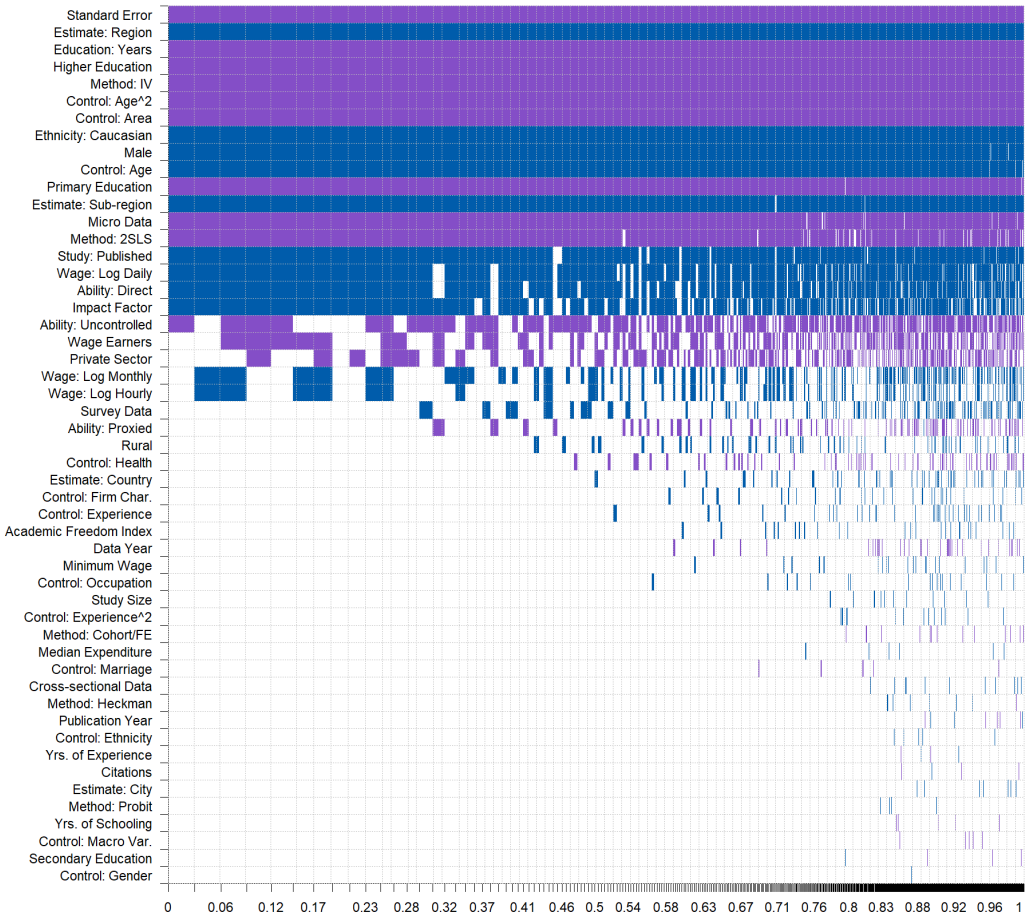
With the large number of variables my dataset holds, it is a rather complex and no doubt challenging task to pick those that could explain the effect behavior the best. Indeed, traditional methods such as OLS are prone to over-specification bias, so simply dumping all collected variables into a single model does not appear like the best approach. Is there a way, then, where we could somehow discern the importance of the collected variables without knowing anything about them a priori? One such technique, called BMA, appears suitable for the task and is precisely what the following sections will focus on.

Pioneered over two decades ago in papers such as Hoeting et al. (1999) or Raftery et al. (1997), the technique provides a balanced approach that considers a multitude of statistically plausible models and assigns different weights to them using the Bayes' theorem and posterior inclusion probabilities. As explained in Hoeting et al. (1999) and Amini & Parmeter (2011), the process then highlights the importance of each variable based on these weights. For the procedures employed in this thesis, it is crucial to understand two metrics - Posterior Model Probability (PMP) and Posterior Inclusion Probability (PIP). For each variable, PMP denotes how well each model fits the data. In contrast, PIP is the sum of posterior model probabilities across the models in which that variable is included. The higher PIP, the higher the variable's importance for explaining the effect's behavior.

I use a combination of the default Zellner's g-prior and the dilution prior for this particular analysis. The choice of the former stems from the fact that this setup allows for more control over collinearity in the data, an issue that may arise with the high number of variables. In my case, the number of eligible individual variables fed into the process once is 52 once reference variables are removed, making the collinearity treatment seem wise. [3] To ensure the carried information is unique for every variable, I also check the variance inflation factors of the included variables. This check revealed the sound design of the dataset, as all 52 variables, when lumped into a single model, hold a variance inflation factor no larger than 10. As for other parts of the setup, the last point

[3]Here, an individual variable refers to each sub-group of a dummy variable group or any other variable that contains multiple categories. For example, the highest achieved education variable, as explained in Subsection 5.1.3, would account for four individual variables (none, primary, secondary, and higher).

Figure 5.1: Bayesian model averaging results



Note: This figure shows the results of the Bayesian model averaging using the uniform g-prior and dilution prior. The response variable, percentage returns to a year of schooling, is measured on the horizontal axis as cumulative posterior model probabilities. The explanatory variables are ranked in descending order on the vertical axis according to their posterior inclusion probability. Purple color: the variable is included in the model and has a positive sign. Blue color: the variable is included in the model and has a negative sign. Numerical results of the estimation can be found in table 5.2. For a detailed explanation of the variables, see table 5.1.

of interest lies perhaps in the choice of the sampler, where I choose to use the default Markov Chain Monte Carlo algorithm (Zeugner & Feldkircher 2015).

As an additional robustness check, I also include FMA with Mallow's criteria for weights (Hansen 2007) and orthogonalization of the model space as per Amini & Parmeter (2012). The reasons for this include higher resiliency against model misspecification or the reduction of model uncertainty. In other words, FMA provides a good sanity check that the BMA setup is not misspecified or overly complex.

I first present graphical results in Figure 5.1. Each variable's contribution

is marked by one of two colors - purple means a positive influence on the effect, while blue represents a negative influence. The columns in the figure each represent a single regression model, while the rows display the inclusion of variables in these models. The left-hand side of the figure shows the best models that best fit the data. The width of each column then captures the individual model's PMP. The proportion of models a variable is included in gives that variable's PIP. For example, if it is included in 50% of models, its PIP will be 0.5. Based on the paper by Kass & Raftery (1995), simple guidelines indicate that PIP values between 0.5 and 0.75 suggest a weak influence on the effect, 0.75-0.9 indicate solid importance of the variable, values over 0.9 and below 0.99 mean strong influence and values over 0.99 are decisive in telling this variable is essential for explaining the effect's behavior. Even glancing into the figure, it is evident that over 15 variables have a PIP over 0.5 in the averaging process. Looking at the fittest model, 19 variables out of 52 are included.

Next, I compare the results for both BMA and FMA, this time quantitatively using numeric coefficients associated with the variables. These are displayed in Table 5.2, where variables with PIP over 0.5 in the BMA have this statistic highlighted. There are 20 variables with PIP over 0.5 in total. When it comes to FMA, p-values of many variables are below 0.001, confirming that the models could identify a large amount of highly important effect drivers.

Looking at these in more detail, the publication bias stands out immediately. Despite the mixed or otherwise lukewarm claims about its presence in Chapter 4, both presented models strongly suggest that publication bias appears in the data. The standard error coefficients are 0.378 and 0.515 for both respective models; both these coefficients are statistically significant. The PIP of 1.000 associated with the coefficient in the BMA model is the highest possible, and the p-value for the FMA model is also below 0.01.

Let us now explore those variables that negatively influence the effect. Firstly, regional and sub-regional data appear to diminish the effect's magnitude, as does reporting the wage in daily units or having a male or Caucasian subject group. Furthermore, published studies are likewise associated with lower effect in both models, although the FMA p-value tied to this claim does not hold enough significance. While the linear age coefficient pulls the effect heavily in the negative direction, the quadratic coefficient works in the opposite direction, ultimately contributing positively to the overall effect. Perhaps the most interesting out of the negative drivers, however, is the direct ability variable. Out of the other ability variables, it is associated with the highest PIP

Table 5.2: Model averaging results

Response variable:	Bayesian model averaging			Frequentist model averaging		
Returns to Year of Schooling	Post. mean	Post. SD	PIP	Coef.	SE	p-value
(Constant)	-4.729	NaN	**1.000**	4.838	350.491	0.989
Standard Error	0.375	0.064	**1.000**	0.516	0.201	0.010
Estimate characteristics						
Estimate: City	-0.006	0.081	0.013	0.000	1.109	0.000
Estimate: Sub-region	-1.479	0.346	**1.000**	-0.612	1.677	0.715
Estimate: Region	-1.334	0.260	**1.000**	-0.699	1.292	0.589
Estimate: Country	-0.030	0.146	0.055	0.000	0.909	0.000
Data Characteristics						
Study Size	-0.002	0.029	0.014	0.000	0.373	0.000
Yrs. of Schooling	0.000	0.003	0.007	0.000	0.003	0.000
Yrs. of Experience	0.000	0.001	0.006	0.000	0.013	0.000
Education: Years	1.149	0.219	**1.000**	1.328	0.619	0.032
Wage: Log Hourly	-0.432	0.465	**0.511**	0.000	0.713	0.000
Wage: Log Daily	-1.611	0.623	**0.963**	-0.595	1.129	0.598
Wage: Log Monthly	-0.671	0.622	**0.602**	0.000	1.011	0.000
Micro Data	1.374	0.309	**0.997**	0.612	0.820	0.455
Survey Data	-0.104	0.238	0.192	0.000	0.584	0.000
Cross-sectional Data	-0.001	0.017	0.005	0.000	0.122	0.000
Data Year	1.172	6.961	0.037	0.000	46.426	0.000
Spatial/structural variation						
Primary Education	3.455	0.855	**0.996**	1.409	2.030	0.488
Secondary Education	-0.003	0.121	0.008	0.000	0.382	0.000
Higher Education	5.397	0.599	**1.000**	4.140	1.514	0.006
Wage Earners	0.882	0.791	**0.621**	0.000	1.411	0.000
Male	-1.202	0.273	**1.000**	-0.657	0.698	0.347
Private Sector	0.800	0.944	0.474	0.000	2.073	0.000
Ethnicity: Caucasian	-1.460	0.258	**1.000**	-1.097	0.546	0.045
Rural	-0.091	0.338	0.083	0.000	1.260	0.000
Median Expenditure	-0.004	0.029	0.032	0.000	0.164	0.000
Minimum Wage	-0.002	0.020	0.021	0.000	0.011	0.000
Academic Freedom Index	-0.018	0.126	0.027	0.000	0.098	0.000
Estimation method						
Method: Cohort/FE	0.005	0.063	0.012	0.000	0.240	0.000
Method: 2SLS	1.529	0.411	**0.996**	0.640	0.989	0.517
Method: Heckman	-0.001	0.031	0.006	0.000	0.063	0.000
Method: Probit	-0.003	0.076	0.008	0.000	0.083	0.000
Method: IV	2.651	0.348	**1.000**	1.701	0.901	0.059
Ability: Direct	-1.218	0.486	**0.930**	-0.632	0.699	0.366
Ability: Proxied	0.085	0.277	0.104	0.000	0.924	0.000
Ability: Uncontrolled	0.492	0.406	**0.696**	0.000	0.950	0.000
Control: Age	-1.921	0.408	**1.000**	-0.983	1.106	0.374
Control: Age^2	2.992	0.432	**1.000**	2.049	1.118	0.067
Control: Experience	-0.021	0.118	0.042	0.000	0.686	0.000
Control: Experience^2	-0.001	0.025	0.007	0.000	0.185	0.000
Control: Ethnicity	0.000	0.022	0.006	0.000	0.206	0.000
Control: Health	0.049	0.188	0.080	0.000	0.600	0.000
Control: Gender	0.000	0.011	0.002	0.000	0.241	0.000
Control: Marriage	0.003	0.038	0.015	0.000	0.254	0.000
Control: Occupation	-0.009	0.079	0.019	0.000	0.005	0.000
Control: Firm Char.	-0.022	0.121	0.045	0.000	0.597	0.000
Control: Area	1.784	0.234	**1.000**	0.840	1.083	0.438
Control: Macro Var.	0.000	0.019	0.007	0.000	0.126	0.000
Publication characteristics						
Impact Factor	-0.215	0.088	**0.931**	-0.105	0.165	0.524
Citations	0.000	0.006	0.006	0.000	0.111	0.000
Study: Published	-1.157	0.280	**0.999**	-0.430	1.242	0.730
Publication Year	0.000	0.017	0.003	0.000	0.044	0.000

Note: This table presents the results of the Bayesian and Frequentist model averaging. Post. mean = Posterior Mean, Post. SD = Posterior Standard Deviation, PIP = Posterior Inclusion Probability, Coef. = Coefficient, SE = Standard Error, OLS = Ordinary Least Squares, FE = Fixed Effects, 2SLS = 2 Stage Least Squares. The variables with PIP > 0.5 are highlighted. For a detailed explanation of the variables, see table 5.1.

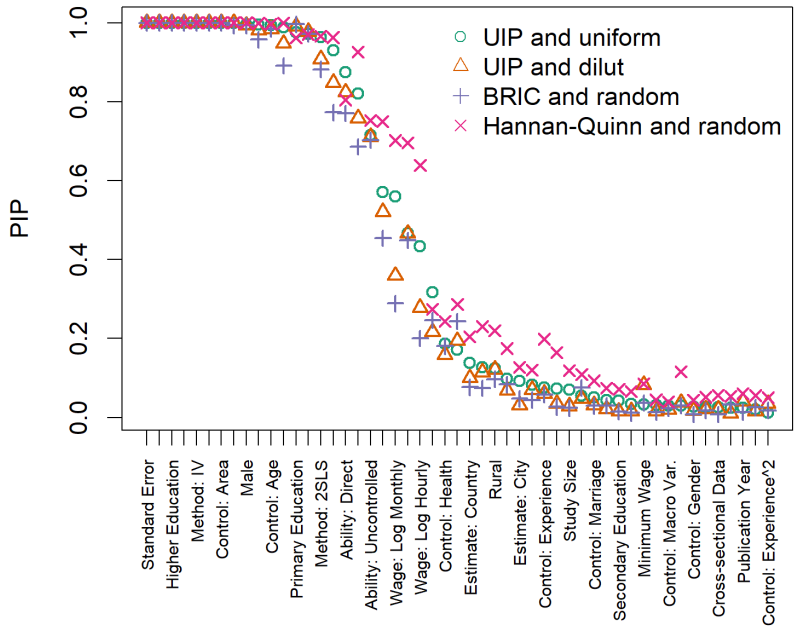
and has the biggest influence on the effect. This potential evidence for ability bias is weakened only by the FMA robustness check, where the p-value is not small enough to claim statistical significance.

Regarding variables that exert a positive influence on the effect as opposed to a negative one, two coefficients stand out the most. These are, first, primary education; second, higher education. Although the secondary education coefficient is insignificant for the BMA model, this may further prove that education truly matters. Apart from this foreseeable conclusion, we can also see that data collected on the micro level positively influence the overall effect, as does estimating the equation using 2-stage least squares or an instrumental variable regression. Controlling for the type of area in which the subjects work also has a significant positive effect, as do the earlier mentioned age squared and standard error. Lastly, I would like to give attention to the *Education: Years* variable, which also exhibits a significant positive impact on the effect. In line with Churchill & Mishra (2018), reporting the estimates in years rather than levels seems to be of systematic importance rather than a fluke. I can think of two sources of this phenomenon - the functional form of Equation 3.2 and human error. While the former may be induced purely by imperfect modeling of the relationship between an attained level of education and the returns associated with each year spent studying for that level, the latter appears less streamlined. Given that all estimates reported in levels had to be transformed and unified using a calculation with incomplete information (sometimes the number of years necessary to finish a certain degree was missing), this uncertainty may give rise to the systematic influence we observe.

As the last piece of information added to the model averaging topic, I also present the differences between posterior inclusion probabilities of models ran under different specifications, namely different priors. Apart from the above-mentioned uniform g-prior and dilution model prior, I run the estimation for three unique pairs of priors. These are, listed in an arbitrary order, uniform g-prior & uniform model prior, benchmark g-prior & random model prior, and Hannah-Quin criterion g-prior & random model prior. The posterior inclusion probabilities of variables ran under all these specifications are displayed in Figure 5.2. The results appear highly stable and invariable towards different specifications, and I find no further need to dig deeper in this regard. For a graphical display of results under each of the three additional specifications, see Appendix B.

This concludes the chapter on model averaging. In Appendix B, you may

Figure 5.2: Inclusion probability varies little across different model specifications



Note: This figure shows how much variables used in Bayesian model averaging contribute to the returns to education effect under different model specifications. The variables are displayed on the x-axis against their posterior inclusion probabilities on the y-axis. PIP = Posterior Inclusion Probability, UIP = Uniform g-prior, Dilut = Dilution Prior, Uniform = Uniform Model Prior, BRIC = Benchmark g-prior, Random = Random Model Prior, HQ = Hannan-Quinn Criterion. For the explanation of the variables and their detailed interpretation, see table 5.1.

find several robustness check figures, including the correlation table of utilized variables, graphical results, and a comparison of BMA models ran under different specifications.

Chapter 6

The best-practice estimate

In this chapter, I would like to focus on one other method that can be used to gain more insight into the effect's behavior. The technique in question involves utilizing the BMA model coefficients from Chapter 5 and actual data values to obtain a best-practice estimate of the effect under different specifications. With this, I hope first to uncover more detail about how different experiment setups change the observed effect and second to bring even more insight into the question of individual variables and the magnitude of their influence on the effect.

6.1 Modelling the best-practice

With the BMA model coefficients from Table 5.2, let us first model a baseline subjective practice by plugging in mostly arbitrary data values. Once this subjective best-practice is obtained, we can then compare it to individual setups of other studies. Regarding the values used for the subjective evaluation, I opt to keep most of them at their mean. It is unclear and, at times, impossible to objectively discern between one and another value of a variable and say which is better. There are, however, two notable exceptions. First, I set the standard error equal to zero, as publication bias is never desirable in the data sample. Second, I utilize the highest available values of the journal impact factor and number of citations available. This stems from the assumption that highly cited studies from top journals should bring more credibility and present estimates close to the true effect.

Apart from this subjective best-practice estimate, I also computed best-practice estimates for all other studies in the dataset. When doing so, an im-

portant question arose of which specifications to use in case a study reported multiple estimates. When a variable remained constant for the study, such as for the number of citations or the impact factor, there was nothing to consider. However, if a variable changed from estimate to estimate, such as when each estimate was associated with a different method, I had to choose a representative value for the whole study. I opted to remedy this by using mode. In other words, I took the value that appeared the most often within the estimates of that study. A weighted mean of observations and their coefficients could have been used instead, but the loss of information incurred by the mode approach should be minimal. Given the large number of variables in the model, it should also not affect the overall picture.

Table 6.1: Implied best-practice

Study	Estimate	95% Confidence Interval	Studies
Author	6.536	(5.762; 7.310)	0
Query	7.529	(3.552; 11.506)	74
Snowballing	6.346	(2.530; 10.162)	41
All studies	7.109	(3.046; 11.17)	115

Note: The table reports estimates of the best-practice estimate according to the author's subjective best-practice, two subsets of the literature, and the whole data sample. For the latter three, the figures are computed by averaging the best-practice estimates of all studies within that data subset. 95% confidence interval bounds are constructed as an approximate using OLS with study level clustered standard errors. Query = Studies identified by query, Snowballing = Studies identified by snowballing, Studies = Number of studies used for the estimation.

As a baseline, I present the results of the implied best-practice calculation using my subjective setup. Then, I display estimates calculated across different subsets of literature for studies identified separately by the query and by the snowballing. Furthermore, I construct an estimate using all studies in the dataset. These can all be found in Table 6.1. The subjective best-practice estimate equals 6.536% with a relatively narrow confidence band. The estimates for the two literature subsets then fall within 1% of the subjective estimate; the query literature predicts 7.529% returns to schooling, while the snowballing literature suggests 6.346%. Understandably, the confidence bounds for these two estimates are much wider, given the large number of studies used in the estimation, together with the fact that this lumps together studies of a much different nature. Still, one could argue that the query studies tend to report higher estimates than their snowballing counterpart, although this claim would lack the statistical significance backup. When looking at the whole dataset, the suggested estimate tallies up to 7.109% with a confidence bound much too wide

to hold any statistical power. Despite this, I believe this number should serve , above all, as a good sanity check, and I think it does just that, given its proximity to the simple literature effect mean identified in Table 3.1.

6.2 Implied best-practice within subsets of literature

To better understand how the implied best-practice behaves within the literature, I calculated how the estimates changed when observed for different data subsets. Using the same variable grouping logic described in Section 6.1, I split the data into an array of subsets and present these in both numeric and graphical formats. I choose this approach over focusing on individual studies as I believe it holds more information, but I append the best-practice estimates for all 115 studies in Appendix C for completeness.

Before getting to the actual results, several points about the technical procedure should be addressed, starting with a point on how the subsetting is done. Different variable types call for a different approach. In my data, I treated these different data types as follows. For dummy variables, the subset consists of studies where that dummy is equal to 0. For variables defined as ratios, such as the ratio of urban vs. rural workers, the subsets include studies where a given variable is the highest out of all its alternatives. For example, suppose that after choosing the most frequent values of the urban vs. rural workers variable, the ratio comes up to 0.25 vs. 0.75 (urban vs. rural). In that case, such a study gets put into the 'rural workers' category, given that these comprise the majority of the sample. The same is true for variables with multiple alternatives, such as for the variable capturing the highest achieved education. Suppose further that the ratio is the same for all variables of the same group. In that case, the representative is chosen randomly, eliminating potentially any bias given a large enough number of studies in the subset. Consequently, results from a sample containing fewer studies should be viewed with caution. And lastly, one note on handling float-type variables. Here, I use the median as the split point and divide the dataset into studies whose representative estimate is above and below this point.

Another important caveat that this approach brings with it lies in using the most frequent value as the representative value for each study. Naturally, this procedure leads to a loss of information, skewing the overall statistics a

bit as a result. For example, the number of citations median of the grouped sample of studies is no longer 80 as in the ungrouped dataset, but 73. As I mentioned in Section 6.1, this could be remedied by weighting each variable with the number of occurrences within the reported estimates of each study. Personally, however, I consider the overall impact of this shortcoming minimal. Moreover, it could be argued that the best practice in literature is the one that each study tends to employ the most, but that, too, is up for debate.

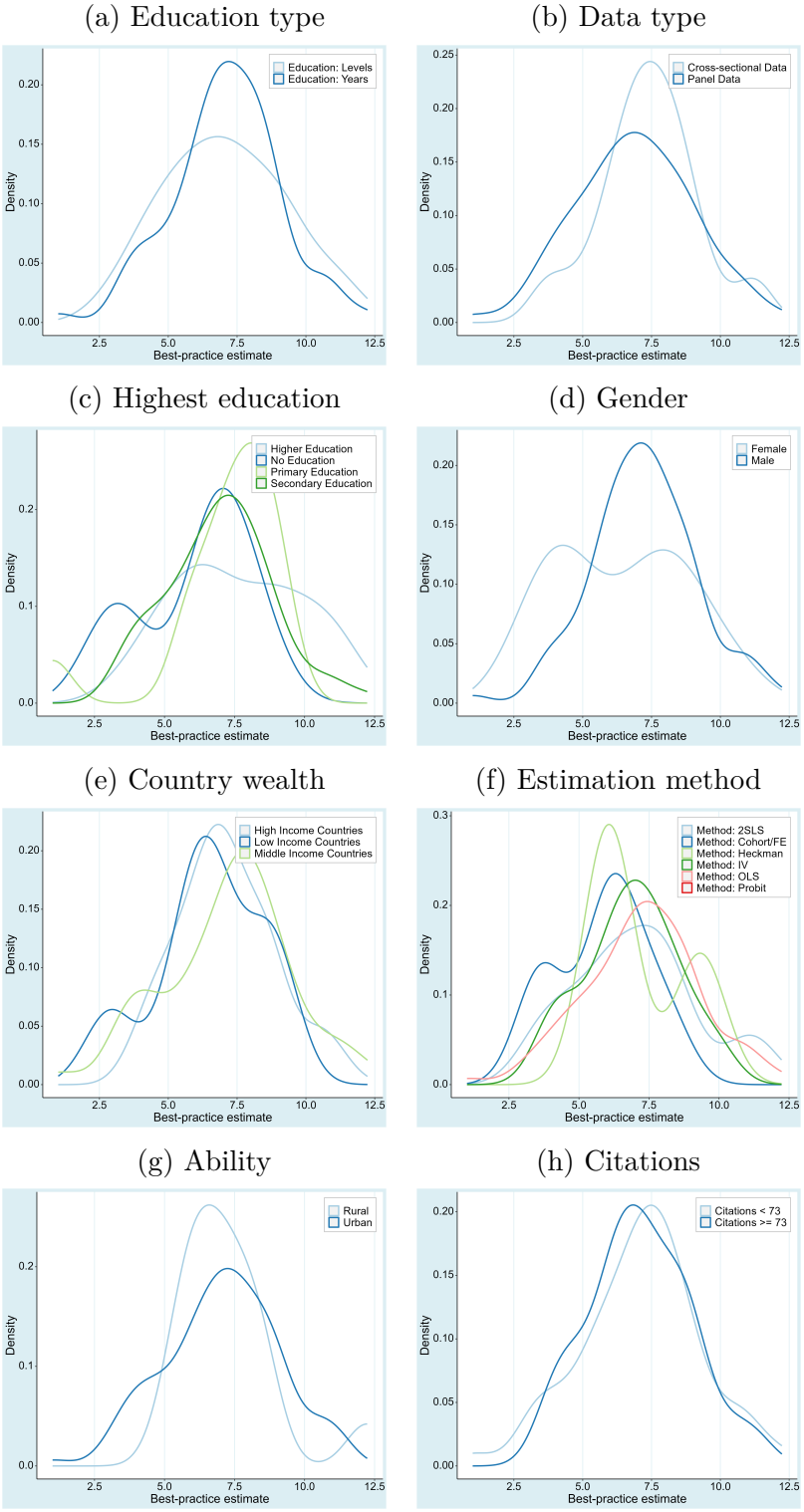
As a last technical point regarding the results, I choose to omit several data subsets from the analysis for presentation clarity. With the formidable number of estimates presented, I firmly believe the shown results accurately represent the overall behavior within the literature and that they paint a clear enough picture. The numeric results can be found in **??**, while the graphs appear in Figure 6.1.

Several crucial points can be drawn from these two sources of information. First and foremost, it is vital to bear in mind the large confidence interval with which nearly all presented estimates and graphs are associated. Despite the lack of confidence bound curves in Figure 6.1, **??** lets us know that the confidence range is quite wide. Second, as I pointed out earlier, the results drawn for data subsets containing only a handful of studies should also be viewed with discretion, as they may be plagued with insufficient data sample bias.

With these considerations in mind, I dare to point out several intriguing patterns within the results. With 7.109% returns to education as a baseline average of all best-practice estimates within the literature, some negative deviations appear for studies focusing on uneducated subjects (6.174%), female subjects (6.388%), studies employing cohort and fixed-effects estimation (5.861%) or the Heckman method (6.417%), and unpublished studies (6.535%). Regarding ability, studies controlling for it directly or with a proxy report on average lower estimates (6.729% and 6.873%, respectively) than studies that omit ability from their models (7.134% and 7.312%). As for variables whose employment in study setup causes an increase in estimates, studies with subjects that attained higher education (7.729%), or were self-employed (8.175%), stand out among the rest. For the most part, even the differences in means are only marginal and amount to only one or two percentage points in returns at most.

In the graphs, the left-hand side of distributions is visibly more prominent for studies focusing on female subjects or controlling for ability in their models, confirming the numeric findings from **??**. A sizable bump of low percentage

Figure 6.1: Implied best-practice across various subsets of data



Note: This figure displays density lines for best-practice estimates of studies employing different variable setups. Each density line corresponds to a subset of studies whose setup involves a particular variable, as described in each graph's legend. The effect of an additional year of schooling on returns is displayed on the x-axis against its density on the y-axis. For Figure 6.1h, the data median is used to determine the subsets. For a description of the variables used in these figures, see Table 5.1.

estimates also appears in the distribution of estimates for studies focusing on uneducated subjects. Still, due to the low number of studies in this subset, it is likely caused by an anomaly in one or two studies' calculations.

6.3 Economic significance

Let us now return to the subjective best-practice estimate and consider the role of individual variables again. Namely, I will calculate the economic significance of some prominent variables, which means observing how much each of these variables contributes to the implied best practice when its value is changed. Variables with low PIP in the model averaging could be argued to have little impact on the effect in the first place, so for this case, I will be considering only those variables that had PIP at least 0.5. In the case of the BMA model outlined in Chapter 5, that totals up to 19 variables. To determine their impact on the effect, I will calculate first how much the implied best-practice changes when there occurs a one standard deviation change in each variable, and then how much that change will be when the variable shifts from its lowest reported value to its highest. The results of these calculations can be found in Table 6.2.

With 7.109% as the reference value of the effect against which the economic significance is compared, there are nine variables with negative influence. In contrast, ten variables pull the effect in the positive direction. Understandably, the standard error is among the variables with a positive sign (0.642 for 1 SD change, 3.435 for maximum change), as an increase in standard error should highly correlate to an increase in the effect. Otherwise, there would have been an unmistakable publication bias in the literature, which the tests in Chapter 4 failed to provide conclusive evidence for. As for the rest of the variables, higher finished education, IV regression, age in the quadratic form, and controlling for area display the highest positive impact among the rest. However, the coefficient for the age squared is offset by its linear counterpart, giving the expected convex shape to the Mincer equation and predicting a substantial increase in earnings later in life. Out of the other variables with positive direction, *Education: Years* stands out the most, underlining the suspicion that estimates reporting education in highest achieved levels instead of in years tend to underestimate the returns to education.

As for the variables with a negative influence on the effect, the most significant change is associated with the aforementioned linear age coefficient. Apart from said coefficient, regional and sub-regional level estimates also diminish

Table 6.2: Economic significance of key variables

	One SD change		Maximum change	
	Effect on Returns	% of BP	Effect on Returns	% of BP
Standard Error	0.635	9.78%	3.399	52.37%
Estimate: Sub-region	-0.442	-6.81%	-1.479	-22.79%
Estimate: Region	-0.616	-9.5%	-1.334	-20.55%
Education: Years	0.554	8.53%	1.149	17.71%
Wage: Log Hourly	-0.216	-3.32%	-0.432	-6.66%
Wage: Log Daily	-0.472	-7.27%	-1.611	-24.81%
Wage: Log Monthly	-0.274	-4.22%	-0.671	-10.34%
Micro Data	0.525	8.09%	1.374	21.17%
Primary Education	0.522	8.04%	3.455	53.23%
Higher Education	1.336	20.58%	5.397	83.14%
Wage Earners	0.181	2.78%	0.882	13.59%
Male	-0.420	-6.48%	-1.202	-18.51%
Ethnicity: Caucasian	-0.612	-9.43%	-1.460	-22.49%
Method: 2SLS	0.449	6.91%	1.529	23.55%
Method: IV	0.832	12.81%	2.651	40.84%
Ability: Direct	-0.416	-6.41%	-1.218	-18.77%
Ability: Uncontrolled	0.243	3.75%	0.492	7.58%
Control: Age	-0.913	-14.07%	-1.921	-29.6%
Control: Age^2	1.336	20.58%	2.992	46.1%
Control: Area	0.880	13.56%	1.784	27.48%
Impact Factor	-0.330	-5.08%	-1.501	-23.13%
Study: Published	-0.491	-7.57%	-1.157	-17.82%

Note: This table shows the individual effect of several important variables on the returns to education, all other things being equal. Variables with PIP at least 0.5 are considered important for this matter. One SD change = Change in the effect incurred by a single standard deviation change in the variable. Maximum change = Change in the effect incurred by increasing the variable from its lowest value to its highest. 6.536% returns to education is the baseline value against which the variables are compared. For an explanation of the variables, see Table 5.1. SD = Standard deviation, 2SLS = Two-stage Least Squares, IV = Instrumental Variable.

the overall effect, as does being a male or Caucasian ethnicity. Further, studies with a high impact factor or studies published in journals tend to report higher estimates than their less recognized counterparts. And last but not least, the issue of ability. As has been the case thus far, controlling for ability directly diminishes the overall effect, while leaving the ability out of the equation is associated with higher returns to education. The size of this ability bias, at least when looking at the economic significance of variables, is relatively smaller. Despite this, its presence is unmistakable and in line with all the results presented thus far.

Chapter 7

Doubling the evidence: Addition of twin studies

So far, I have explored the role of schooling and its contribution to an individual's future earnings. Furthermore, I tried to answer the question of what role ability plays in this equation and whether or not it should be accounted for. Even though I claim that some magnitude of ability bias exists in the relationship, one crucial question remains unanswered. One that, I deemed, deserves an extra chapter of attention, and the ignoring of which may take away credibility from the presented results. This question is - to what extent is the increase in earnings influenced by schooling and to what extent by ability? Is there a way to separate these two and isolate the effect schooling has on an individual's wage, regardless of their ability? As it turns out, there is. Using a sample of twins, one may theoretically rule out the role of ability and family background and observe the unbiased influence of education on earnings. In the following chapter, I attempt to take this approach by constructing an entirely new dataset containing only twin studies. With this dataset, I will run the analysis anew and try to determine whether individual differences and innate ability play a crucial role in determining one's future or whether it is all just a matter of education.

7.1 Understanding natural experiments: Is it all intertwined?

For this analysis, it is vital to understand how being a twin plays a significant role in the matter. We can identify two types of twins - monozygotic and

dizygotic. Monozygotic twins (marked further as MZ twins), sometimes called identical twins, come from a single zygote and thus share the same genetic information. For this, it is reasonable to assume they share the same innate ability, and any noteworthy differences that arise during their lifetime should come from their environment, schooling, family background, etc. Dizygotic twins (marked further as DZ twins), on the other hand, come from two different zygotes, and their genetic information thus differs slightly. As such, these may be looked at more as siblings of the same age. Using this crucial difference will allow us to compare the two and further single out the role of innate ability even further. In both cases, however, the role of family background is a common factor that may be observed and compared with the population sample, where such similarities are not present.

Now let us take a look at the existing literature on the topic. Like twins sharing the same family background and living environment, the existing studies also share many similarities. The two studies of Ashenfelter & Krueger (1994) and Ashenfelter & Rouse (1998) perhaps stand as the cornerstones behind the current form and direction of the literature. In these studies, the authors construct a survey and study samples of twins to find that the OLS estimate of returns to schooling is upward biased. In the Ashenfelter & Krueger (1994), the authors propose a new approach where the measurements of education are collected from both twins, and the final estimate comes from the within-twin comparison. This involves taking one twin's report of the within-twin schooling as an instrument for the other twin's report. The benefit of this approach lies in addressing the possible measurement error that sometimes arises in schooling reports. Several other studies, including Behrman et al. (1994), Isacsson (1999), or Bonjour et al. (2003), too follow a similar approach and provide a solid theoretical background to the matter.

Another vital issue, as well as a critique of the twin approach, lies in the idea that the within-twin schooling differences may not be random but endogenous with respect to wages. Bound & Solon (1999), for example, argue that ability can be influenced by factors other than genes and that using methods such as IV regression to remedy the measurement error can simultaneously increase the omitted ability bias. On the other hand, using techniques such as Fixed-effects estimator may remove the omitted variable bias but does so at the cost of introducing even greater bias in measurement error (Ning 2005). Despite this, both suggested methods take care of at least one issue of the approach and, when combined, may serve to paint a clearer picture of the overall effect.

Since this thesis is primarily concerned with the role of ability, I opt for the following approach. I define six methods that all focus on fixing different issues - OLS, Generalized Least Squares (GLS), IV regression, Fixed-effect estimator, First-differences, and First-differences by IV. Regarding ability, the former three are more likely to be plagued by the omitted ability bias, while the latter three do consider it and should account for it. By comparison of these two groups, it should be possible to identify the magnitude of the omitted variable effect in the data (Rouse 1999). In this thesis, I will not focus on the influence of the measurement error.

7.2 What do you mean there are two?: Making a twin dataset

I will construct the new dataset, comprising natural experiments, with two analysis goals in mind. First, as described in the previous section, I will attempt to quantify the omitted variable bias, and second, I will want to compare the results with the conclusions obtained from the earlier chapters. As such, the form of the dataset will be nearly identical to the one described in Chapter 3, with slight modifications to accommodate the specific design of the included studies.

As for the studies themselves, I start with the literature review of Nakamuro & Inui (2012) and Li et al. (2012), and from there, perform snowballing to identify as many relevant studies on the topic as possible. Using this approach, I collected, in total, 16 studies. For their complete list, see Appendix A. Given how intertwined the studies on the topic are, perhaps due to the relatively small scope of the topic, the choice of which papers to include was somewhat streamlined. Possibly, I may have missed several studies, but I am highly confident that this set should provide a highly representative sample of the literature.

After collecting the data from these 16 studies, it became apparent that some variables were unusable for this particular use case. Two variables, *Sector: Public/Private*, *Sector: Urban/Rural*, had no observations associated with them at all, while the variables *Control: Experience squared*, *Control: Occupation*, and *Education: Primary/Secondary/...* had fewer than ten. As such, I removed all these variables from the dataset, together with the *Ability* variable, for the approach to measuring ability bias is slightly different now, as

explained in Section 7.1. For some variable groups, only some sub-categories had no data, such as *Estimate: Sub-region/Continent*, *Micro Data*, *Region: Lat-America/Middle East and North Africa/South Asia/Saharan Africa*, *Income: Low*, *Instrument: Distance to school*, and finally, *Control: Health*. On the other hand, I also added a handful of new variables, including:

- *White/Non-white* - Ratio of white subjects to non-white subjects.
- *Married/Unmarried* - Ratio of married subjects to non-married subjects.
- *Identical/Non-identical/No twins* - Ratio of subjects that are either identical (MZ) or non-identical (DZ) twins or are not twins at all.
- *Method: Selection/FE* - =1 if the authors use Selection-effects or Fixed-effects estimation.
- *Method: IV First-differenced* - =1 if the authors use First-Differenced IV estimation.
- *Instrument: Smoking* - =1 if the authors use smoking as an instrument in the regression.

For the list of all variables used in the analysis and their descriptive statistics, see Table 7.1. For the list of descriptions of the rest of the variables, see Table 5.1. The final form of the new dataset includes 293 observations across 16 studies and can be found in the online appendix.

For brevity's sake, I choose not to focus in depth during the analysis on differences between subsets of data, save for the type of method used. However, several statistics that characterize the new group of subjects might be helpful to highlight here just to get a better picture of how this new dataset differs from the old one. Firstly, two-thirds of the data consist of identical twins, roughly 26% of non-identical twins, and less than 10% of non-twin subjects. Among these, about 70% are married, nearly the same amount are white, and over 82% live in high-income countries. The subjects spent, on average, 12.463 in school and 17.875 years working. For over 95% of them, the schooling statistic is reported in years, as opposed to levels. Other statistics, including variable groups capturing data type, estimation method, publication characteristics, etc., can all be found in the aforementioned Table 7.1.

Table 7.1: Variables of the twin dataset

Variable	Mean	SD	Obs	Variable	Mean	SD	Obs
Effect	6.251	2.76	293	Income: High	0.823	0.383	241
Estimate characteristics				Income: Middle	0.177	0.383	52
Standard Error	1.126	1.036	293	Median Expenditure	4.214	3.901	293
Estimate: City	0.338	0.474	99	Minimum Wage	3.025	2.761	293
Estimate: Region	0.055	0.228	16	Acad. Freedom Index	0.786	0.241	293
Estimate: Country	0.608	0.489	178	Mean Age	3.653	0.100	293
Data characteristics				*Estimation method*			
Study Size	3.000	0.426	293	Method: OLS	0.345	0.476	101
Yrs. of Schooling	12.463	1.456	293	Method: GLS	0.102	0.304	30
Yrs. of Experience	17.875	5.627	293	Method: Selection/FE	0.253	0.435	74
Education: Years	0.959	0.199	281	Method: FD	0.034	0.182	10
Education: Levels	0.041	0.199	12	Method: IV-FD	0.068	0.253	20
Wage: Hourly	0.287	0.453	84	Method: IV	0.198	0.399	58
Wage: Daily	0.130	0.337	38	Instr.: Sibling Ed.	0.140	0.348	41
Wage: Monthly/Annual	0.584	0.494	171	Instr.: Smoking	0.061	0.241	18
Survey Data	0.689	0.464	202	Instr.: Other	0.048	0.214	14
National Register Data	0.311	0.464	91	Control: Age	0.584	0.494	171
Cross-sectional Data	0.498	0.501	146	Control: Age^2	0.478	0.5	140
Panel Data	0.502	0.501	147	Control: Experience	0.218	0.414	64
Data Year	3.297	1.286	293	Control: Ethnicity	0.157	0.364	46
Spatial/Structural variation				Control: Gender	0.522	0.5	153
Wage Earners	0.962	0.052	102	Control: Marriage	0.416	0.494	122
Gender: Male	0.557	0.277	265	Control: Firm Char.	0.123	0.329	36
Gender: Female	0.443	0.277	28	Control: Area	0.055	0.228	16
White	0.694	0.421	190	Control: Macro Var.	0.038	0.19	11
Ethnicity: Caucasian	0.208	0.407	61	*Publication characteristics*			
Married	0.699	0.143	268	Impact Factor	-0.269	1.125	206
Unmarried	0.301	0.143	252	Citations	3.855	2.173	293
Twins: Identical	0.640	0.415	242	Study: Published	0.703	0.458	206
Twins: Non-Identical	0.263	0.377	126	Study: Unpublished	0.297	0.458	87
Twins: None	0.097	0.275	34	Publication Year	1.018	0.898	264

Note: This table presents basic summary statistics for variables of the new twin dataset. For detailed descriptions of all variables unmentioned in this chapter, see Table 5.1. SD = Standard Deviation, OLS = Ordinary Least Squares, GLS = Generalized Least Squares, FE = Fixed-Effects, IV = Instrumental Variable.

7.3 Empirical analysis: Are the results just identical?

As far as the outcome of the analysis is concerned, I will focus only on a handful of easily presentable results to keep the chapter concise. The first is the publication bias issue, which can be summarized in a single table and should not be omitted even from this analysis. In Table 7.2, I present the results of all linear, non-linear, and endogeneity-robust tests and methods explained in Chapter 4. That is, all but one, as there were too many non-linear techniques to fit into the table nicely; I decided to remove the Hierarchical Bayes results arbitrarily. For clarity, I add that the test for publication bias using said method yielded a coefficient of 0.6 with a standard error of 0.365. In contrast, the coefficient associated with the effect was estimated at 6.857 with a standard error of 0.544 over 293 observations. The rest of the results can be found in the table mentioned above.

A clear takeaway from these tests, which also holds across different methods and approaches, is that publication bias is considerably more prominent in the twin dataset than in the primary dataset, as explored in Chapter 4. With 6.267% from Table 7.1 as the baseline, the returns to schooling drop by an average of two, sometimes up to three percentage points. Namely, the STEM-based method suggests returns to education of 3.403%, while the Endogenous Kink approach claims 3.908%. On the other end of the spectrum, WAAP and the Selection model report the highest returns, 5.77%, and 5.616%, respectively. No method reports a coefficient of schooling higher than the simple data average, save for p-uniform*, where the standard error failed to be estimated.

Moving on from publication bias, the crux of the matter, which I would like to focus on in terms of analysis outcomes, is the influence of different methods. As outlined in Section 7.1, the difference in approach to the omitted ability bias could have us conclude that if the results of the six employed methods vary only a little, the ability bias in the twin studies is not present. On the other hand, if they vary greatly, meaning the estimates of methods controlling for ability are lower than their counterparts, then that would suggest evidence to the contrary.

As a first insight into the influence of different methods on the outcome, I present both numerically and graphically how the returns to education effects behave when employing each of the six methods. In Figure 7.1, densities of

Table 7.2: Twin studies are plagued by publication bias

Panel A: Linear methods	**OLS**	**FE**	**RE**	**Study**	**Precision**
Publication bias	1.347***	0.602***	0.840***	0.947***	2.897***
(Standard error)	(0.138)	(0.162)	(0.154)	(0.177)	(0.442)
Effect beyond bias	4.735***	5.574***	5.55***	4.754***	3.907***
(Constant)	(0.175)	(0.219)	(0.342)	(0.185)	(0.232)
Observations	293	293	293	293	293
Panel B: Non-linear methods	**WAAP**	**Top10**	**Stem**	**AK**	**Kink**
Publication bias				2.257***	2.895***
				(0.126)	(0.435)
Effect beyond bias	5.77***	4.314***	3.403***	5.616***	3.908***
	(0.159)	(0.265)	(0.95)	(0.157)	(0.093)
Observations	293	293	293	293	293
Panel C: Methods relaxing the exogeneity assumption				**IV**	**p-uniform***
Publication bias				1.824***	L = 1.712
				(0.159)	(p = 0.191)
Effect beyond bias				4.198***	7.79
				(0.188)	(NA)
Observations				293	293

Note: Panel A: Results obtained from estimating the linear equation Equation 4.1. Standard errors, clustered at the study level, are included in parentheses. OLS = Ordinary Least Squares. FE = Fixed Effects. RE = Random Effects. Precision = Estimates are weighted by the inverse of their standard error. Study = Estimates are weighted by the inverse number of observations reported per study. Panel B: Estimates of the effect and publication bias using five non-linear methods. WAAP = Weighted Average of the Adequately Powered (Ioannidis et al. 2017), Top10 = Top10 method by Stanley et al. (2010), Stem = the stem-based method by Furukawa (2019) where P represents the probability of results insignificant at 5% are published relative to the probability of the significant ones at the same level, AK = Andrews & Kasy (2019)'s Selection model, Kink = Endogenous kink model by Bom & Rachinger (2019). Standard errors, clustered at the study level, are included in parentheses. Panel C: Estimates of the effect and publication bias using two techniques that relax the exogeneity assumption. IV = Instrumental Variable Regression; the inverse of the square root of the number of observations is used as an instrument for the standard error. Standard errors, reported in parentheses, are also clustered at the study level. P-uniform* = method proposed by van Aert & van Assen (2021); L represents the publication bias test t-statistic, the corresponding p-value can be found in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 7.3: Summary statistics for the twin dataset using different estimation methods

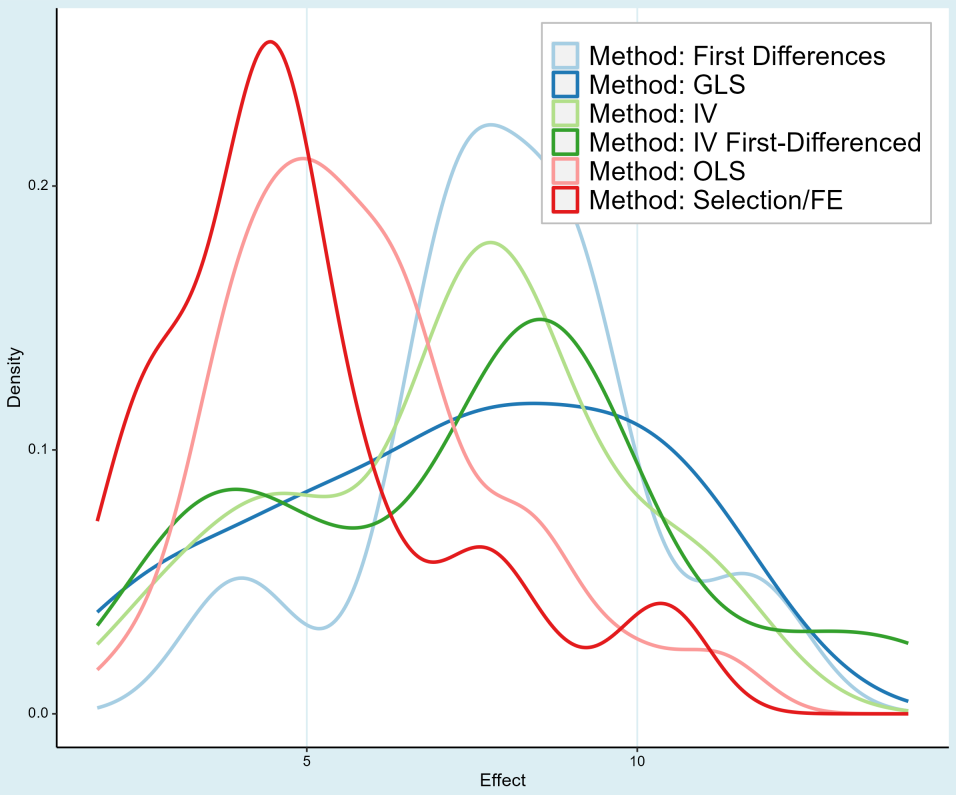
	Unweighted			Weighted			
	Mean	95% conf. int.		Mean	95% conf. int.		N. obs
Baseline methods							
Method: OLS	5.686	1.648	9.724	5.754	1.716	9.792	101
Method: GLS	7.363	1.822	12.904	8.005	2.464	13.546	30
Method: IV	7.155	1.644	12.666	7.570	2.059	13.081	58
Methods that treat the omitted ability bias							
Method: Selection/FE	4.917	0.515	9.319	5.630	1.228	10.032	74
Method: First Differences	7.920	3.979	11.861	7.916	3.975	11.857	10
Method: IV First-Differenced	8.689	1.035	16.343	8.725	1.071	16.379	20

Note: This table presents basic summary statistics of the returns to an additional year of schooling coefficient calculated on various subsets of the data. Unweighted = Original dataset is used. Weighted = Estimates are weighted by the inverse number of estimates reported by each study. OLS = Ordinary Least Squares, GLS = Generalised Least Squares, IV = Instrumental Variable, FE = Fixed-Effects. For cutoff points, medians are used except for dummy variables, where the cutoffs are 0.5.

the effect are displayed under different method specifications, while descriptive statistics of the effect behavior under said specifications can be found in Table 7.3.

No obvious pattern appears between the two groups of methods, that is, between the group that does treat the omitted ability bias and the one that does not. While estimates reported by OLS and the Selection/Fixed-Effects methods suggest the lowest estimates of all (5.686% and 4.917%, respectively), when observing the difference in their estimates as per Li et al. (2012), it comes up to only 0.769% (5.686 - 4.917). Even when looking at the weighted average of estimates for these methods, the discrepancies are overall minimal. On balance, this simple glance into the data suggests that treating the ability during the estimation of twin data samples presents only a marginal effect.

Figure 7.1: Returns to education for twins vary based on the method



Note: This figure displays the densities of the effect of returns to education for twins under different method specifications. The effect is plotted on the x-axis against its density on the y-axis. For a description of the variables used in these figures, refer to Section 7.2.

Chapter 8

Conclusion

Text

Bibliography

Aakvik, A., Salvanes, K. G., & Vaage, K. (2010). Measuring heterogeneity in the returns to education using an education reform. *European Economic Review*, *54*(4), 483–500.

Acemoglu, D. & Angrist, J. (1999). How large are the social returns to education? Evidence from compulsory schooling laws.

Agrawal, T. (2012). Returns to education in india: Some recent evidence.

Allenby, G. M. & Rossi, P. E. (2006). Hierarchical bayes models.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, volume 4 (pp. 1–181). Elsevier.

Amini, S. M. & Parmeter, C. F. (2011). Bayesian model averaging in r. *Journal of Economic and Social Measurement*, *36*(4), 253–287.

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

Andrews, I. & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–94.

Angrist, J. D. (1995). The economic returns to schooling in the west bank and gaza strip. *American Economic Review*, *85*(5), 1065–1087.

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arkes, J. (2010). Using unemployment rates as instruments to estimate returns to schooling. *Southern Economic Journal*, *76*(3), 711–722.

Aromolaran, A. B. (2006). Estimates of mincerian returns to schooling in nigeria. *Oxford Development Studies*, *34*(2), 265–292.

Aryal, G., Bhuller, M., & Lange, F. (2022). Signaling and employer learning with instruments. *American Economic Review*, *112*(5), 1669–1702.

Asadullah, M. N. (2006). Returns to education in bangladesh. *Education Economics*, *14*(4), 453–468.

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

Ashenfelter, O. & Krueger, A. (1994). Estimates of the economic return to schooling from a new sample of twins.

Ashenfelter, O. & Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics*, *113*(1), 253–284.

Ashenfelter, O. C. & Rouse, C. E. (1999). Schooling, intelligence, and income in america: Cracks in the bell curve.

Aslam, M. (2007). Rates of return to education by gender in pakistan. *Education Economics*, *15*(2), 209–224.

Ayyash, M., Sadeq, T., & Sek, S. K. (2020). Returns to schooling in palestine: a bayesian approach. *International Journal of Education Economics and Development*, *11*(1), 37–57.

Bakis, O., Davutyan, N., Levent, H., & Polat, S. (2013). Quantile estimates for social returns to education in turkey: 2006–2009. *Middle East Development Journal*, *5*(3), 1350017–1.

Bartolj, T., Ahčan, A., Feldin, A., & Polanec, S. (2013). Evolution of private returns to tertiary education during transition: evidence from slovenia. *Post-Communist Economies*, *25*(3), 407–424.

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. (2023). Robust bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, *14*(1), 99–116.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9–49.

Behrman, J. R. & Rosenzweig, M. R. (1999). â€śabilityâ€t biases in schooling returns and twins: a test and new estimates. *Economics of education review*, *18*(2), 159–167.

Behrman, J. R., Rosenzweig, M. R., & Taubman, P. (1994). Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment. *Journal of political economy*, *102*(6), 1131–1174.

Belzil, C. & Hansen, J. (2002). Unobserved ability and the return to schooling. *Econometrica*, *70*(5), 2075–2091.

Belzil, C. & Hansen, J. (2004). Earnings dispersion, risk aversion and education. In *Accounting for worker well-being*, volume 23 (pp. 335–358). Emerald Group Publishing Limited.

Bergman, E. & Schöön, C.-G. (2018). Returns to schooling and potential signalling effects: Estimates based on issp data on sweden. *Research in Social Stratification and Mobility*, *56*, 54–67.

Berman, E., Lang, K., & Siniver, E. (2003). Language-skill complementarity: returns to immigrant language acquisition. *Labour Economics*, *10*(3), 265–290.

Bingley, P., Christensen, K., & Walker, I. (2009). The returns to observed and unobserved skills over time: Evidence from a panel of the population of danish twins.

Blackburn, M. L. & Neumark, D. (1993). Are ols estimates of the return to schooling biased downward? another look.

Blanchflower, D. & Elias, P. (1999). Ability, schooling and earnings: Are twins different?

Blundell, R., Dearden, L., & Sianesi, B. (2001). Estimating the returns to education: Models, methods and results.

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

Bonjour, D., Cherkas, L. F., Haskel, J. E., Hawkes, D. D., & Spector, T. D. (2003). Returns to education: Evidence from uk twins. *American Economic Review*, *93*(5), 1799–1812.

Botchorishvili, V. (2007). Private returns to education in georgia. Technical report, Economics Education and Research Consortium, National University, Kyir-Mohyla Academy.

Bound, J. & Solon, G. (1999). Double trouble: on the value of twins-based estimation of the return to schooling. *Economics of Education Review*, *18*(2), 169–182.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of economic literature*, *39*(4), 1137–1176.

Brainerd, E. (1998). Winners and losers in russia's economic transition. *American Economic Review*, *88*(5), 1094–1116.

Breda, T. (2014). Firms' rents, workers' bargaining power and the union wage premium. *The Economic Journal*, *125*(589), 1616–1652.

Campaniello, N., Gray, R., & Mastrobuoni, G. (2016). Returns to education in criminal organizations: Did going to college help michael corleone? *Economics of Education Review*, *54*, 242–258.

Campos, M. & Reis, H. (2017). Revisiting the returns to schooling in the portuguese economy. *Banco de Portugal Economic Studies*, *3*(2), 1–28.

Capatina, E. (2014). Skills and the evolution of wage inequality. *Labour Economics*, *28*(C), 41–57.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of labor market behaviour: Essays in honour of John Vanderkamp* (pp. 201–222). University of Toronto Press.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, *3*, 1801–1863.

Card, D. & Krueger, A. B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *Journal of political Economy*, *100*(1), 1–40.

Carneiro, P., Heckman, J. J., & Vytlacil, E. (2011). Estimating marginal returns to education. *The American Economic Review*, *101*(6), 2754–2781.

Casado-Díaz, J. M. & Lillo-Bañuls, A. (2005). How profitable is to study in spain? an empirical insight using a new source of information.

Chanis, S., Eleftheriou, K., Hadjidema, S., & Katavelis, V. (2021). Tell me who your co-worker is and i will tell you how much you earn: human capital spillovers in the greek health sector. *Journal of Education and Work*, *34*(2), 128–142.

Chase, R. S. (1998). Markets for communist human capital: Returns to education and experience in the czech republic and slovakia. *ILR Review*, *51*(3), 401–423.

Churchill, S. A. & Mishra, V. (2018). Returns to education in china: a meta-analysis. *Applied Economics*, *50*(54), 5903–5919.

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., Mcllroy, W., & Oxman, A. D. (1993). Should unpublished data be included in meta-analyses?: Current convictions and controversies. *Jama*, *269*(21), 2749–2753.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Bernhard, M., Cornell, A., Fish, M. S., Gastaldi, L., Gjerløw, H., Glynn, A., Good God, A., Grahn, S., Hicken, A., Kinzelbach, K., Krusell, J., Marquardt, K. L., McMann, K., Mechkova, V., Medzihorsky, J., Paxton, P., Pemstein, D., Pernes, J., Rydén, O., von Römer, J., Seim, B., Sigman, R., Skaaning, S.-E., Staton, J., Sundström, A., Tzelgov, E., Wang, Y.-t., Wig, T., Wilson, S., & Ziblatt, D. (2023). V-dem country-year dataset v13.

Cui, Y. & Martins, P. S. (2021). What drives social returns to education? a meta-analysis. *World Development*, *148*, 105651.

De Brauw, A. & Rozelle, S. (2008). Reconciling the returns to education in off-farm wage employment in rural china. *Review of Development Economics*, *12*(1), 57–71.

Deary, I. J. (2020). *Intelligence: A very short introduction*, volume 39. Oxford University Press, USA.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21.

Depken, C., Chiseni, C., & Ita, E. (2019). Returns to education in south africa: Evidence from the national income dynamics study. *Zagreb International Review of Economics & Business*, *22*(1), 1–12.

Devereux, P. & Hart, R. (2010). Forced to be rich? returns to compulsory schooling in britain. *Economic Journal*, *120*(549), 1345–1364.

Doan, T., Strazdins, L., & Leach, L. (2020). Cost of poor health to the labour market returns to education in australia: another pathway for socio-economic inequality. *The European Journal of Health Economics*, *21*, 635–648.

Dougherty, C. R. & Jimenez, E. (1991). The specification of earnings functions: tests and implications. *Economics of Education Review*, *10*(2), 85–98.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, *91*(4), 795–813.

Dumauli, M. T. (2015). Estimate of the private return on education in indonesia: Evidence from sibling data. *International Journal of Educational Development*, *42*, 14–24.

Duraisamy, P. (2002). Changes in returns to education in india, 1983–94: by gender, age-cohort and location. *Economics of Education Review*, *21*(6), 609–622.

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

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

Fang, H., Eggleston, K. N., Rizzo, J. A., Rozelle, S., & Zeckhauser, R. J. (2012). The returns to education in china: Evidence from the 1986 compulsory education law. Technical Report w18189, National Bureau of Economic Research.

Feigenbaum, J. J. & Tan, H. R. (2020). The return to education in the mid-twentieth century: Evidence from twins. *The Journal of Economic History*, *80*(4), 1101–1142.

Fersterer, J., Pischke, J.-S., & Winter-Ebmer, R. (2008). Returns to apprenticeship training in austria: Evidence from failed firms. *The Scandinavian Journal of Economics*, *110*(4), 733–753.

Fleisher, B. M., Sabirianova, K., & Wang, X. (2005). Returns to skills and the speed of reforms: Evidence from central and eastern europe, china, and russia. *Journal of comparative economics*, *33*(2), 351–370.

Fortin, N. (2008). The gender wage gap among young adults in the united states: The importance of money versus people. *The Journal of Human Resources*, *43*(4), 884–918.

Fox, J. & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Frazer, G. (2023). Firm productivity, worker ability, and returns to education. *Econometrica*, *91*(1), 107–135.

Furukawa, C. (2019). Publication bias under aggregation frictions: Theory, evidence, and a new correction method.

Gerber, A., Malhotra, N., et al. (2008). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*(3), 313–326.

Gibson, J. & Fatai, O. K. (2006). Subsidies, selectivity and the returns to education in urban papua new guinea. *Economics of Education Review*, *25*(2), 133–146.

Giles, J., Park, A., & Wang, M. (2019). The great proletarian cultural revolution, disruptions to education, and the returns to schooling in urban china. *Economic Development and Cultural Change*, *68*(1), 131–164.

Gill, A. M. & Leigh, D. E. (2000). Community college enrollment, college major, and the gender wage gap. *ILR Review*, *54*(1), 161–181.

Girma, S. & Kedir, A. (2005). Heterogeneity in returns to schooling: Econometric evidence from ethiopia. *The Journal of Development Studies*, *41*(8), 1405–1416.

Glewwe, P. (1996). The relevance of standard estimates of rates of return to schooling for education policy: A critical assessment. *Journal of Development economics*, *51*(2), 267–290.

Gorodnichenko, Y. & Peter, K. S. (2005). Returns to schooling in russia and ukraine: A semiparametric approach to cross-country comparative analysis. *Journal of Comparative Economics*, *33*(2), 324–350.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems.

Grogger, J. & Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *The Journal of Human Resources*, *30*(2), 280–310.

Guifu, C. & Hamori, S. (2009). Economic returns to schooling in urban china: Ols and the instrumental variables approach. *China Economic Review*, *20*(2), 143–152.

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

Harmon, C., Oosterbeek, H., & Walker, I. (2002). The returns to education: A review of evidence, issues and deficiencies in the literature.

Harmon, C., Oosterbeek, H., & Walker, I. (2003). The returns to education: Microeconomics. *Journal of Economic Surveys*, *17*(2), 115–141.

Harmon, C. & Walker, I. (1995). Estimates of the economic return to schooling for the united kingdom. *American Economic Review*, *85*(5), 1278–1286.

Harmon, C. & Walker, I. (1999). The marginal and average returns to schooling in the uk. *European Economic Review*, *43*, 879–887.

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

Hawley, J. D. (2004). Changing returns to education in times of prosperity and crisis, thailand 1985–1998. *Economics of Education Review*, *23*(3), 273–286.

Heckman, J. & Polachek, S. (1974). Empirical evidence on the functional form of the earnings-schooling relationship. *Journal of the American Statistical association*, *69*(346), 350–354.

Heckman, J. & Vytlacil, E. (2001). Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. *Review of Economics and Statistics*, *83*(1), 1–12.

Heckman, J. J. (1979). Sample selection bias as a specification error.

Heckman, J. J., Lochner, L., & Todd, P. E. (2003). Fifty years of mincer earnings regressions.

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. In *Handbook of the Economics of Education (Volume 1)* (pp. 307–458). Elsevier.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the highscope perry preschool program. *Journal of public Economics*, *94*(1-2), 114–128.

Heckman, J. J. & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review*, *91*(2), 145–149.

Herrnstein, R. J. & Murray, C. (2010). *The bell curve: Intelligence and class structure in American life*. Simon and Schuster.

Himaz, R. & Aturupane, H. (2016). Returns to education in sri lanka: a pseudo-panel approach. *Education Economics*, *24*(3), 300–311.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, *14*(4), 382–417.

Horie, N. & Iwasaki, I. (2023). Returns to schooling in european emerging markets: a meta-analysis. *Education Economics*, *31*(1), 102–128.

Hubbard, W. H. J. (2011). The phantom gender difference in the college wage premium. *The Journal of Human Resources*, *46*(3), 568–586.

Ichino, A. & Winter-Ebmer, R. (1999). Lower and upper bounds of returns to schooling: An exercise in iv estimation with different instruments. *European Economic Review*, *43*, 889–901.

Ichino, A. & Winter-Ebmer, R. (2004). The long-run educational cost of world war ii. *Journal of Labor Economics*, *22*(1), 57–87.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research.

Irsova, Z., Bom, P. R., Havranek, T., & Rachinger, H. (2023). Spurious precision in meta-analysis.

Isacsson, G. (1999). Estimates of the return to schooling in sweden from a large sample of twins. *Labour Economics*, *6*(4), 471–489.

Isacsson, G. (2004). Estimating the economic return to educational levels using data on twins. *Journal of Applied Econometrics*, *19*(1), 99–119.

Iwasaki, I. & Ma, X. (2021). Returns to secondary and tertiary education in china: A meta-analysis. *Asian Economics Letters*, *3*(1).

Jones, P. (2001). Are educated workers really more productive? *Journal of Development Economics*, *64*, 67–79.

Joseph, C. (2020). Education and labour market earnings in low income countries: Empirical evidence for tanzania. *Tanzania Journal for Population studies and Development*, *26*(2).

Kane, T. J. & Rouse, C. E. (1993). Labor market returns to two- and four-year colleges: Is a credit a credit and do degrees matter? Technical report, National Bureau of Economic Research.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kenayathulla, H. B. (2013). Higher levels of education for higher private returns: New evidence from malaysia. *International Journal of Educational Development*, *33*(4), 380–393.

Kijima, Y. (2006). Why did wage inequality increase? evidence from urban india 1983–99. *Journal of Development Economics*, *81*(1), 97–117.

Kingdon, G. G. (1998). Does the labour market explain lower female schooling in india? *Journal of Development Studies*, *35*(1), 39–65.

Kolstad, I. & Wiig, A. (2015). Education and entrepreneurial success. *Small Business Economics*, *44*, 783–796.

Krafft, C. (2018). Is school the best route to skills? returns to vocational school and vocational skills in egypt. *The Journal of Development Studies*, *54*(7), 1100–1120.

Krafft, C., Branson, Z., & Flak, T. (2019). What's the value of a degree? evidence from egypt, jordan and tunisia. *Compare: A Journal of Comparative and International Education*, *49*(5), 784–806.

Leigh, A. (2008). Returns to education in australia. *Economic Papers: A journal of applied economics and policy*, *27*(3), 233–249.

Leigh, A. & Ryan, C. (2008). Estimating returns to education using different natural experiment techniques. *Economics of Education Review*, *27*, 149–160.

Lemieux, T. & Card, D. (2001). Education, earnings, and the â€ścanadian g.i. billâ€ť. *The Canadian Journal of Economics*, *34*(2), 313–344.

Li, H., Liu, P. W., & Zhang, J. (2012). Estimating returns to education using twins in urban china. *Journal of Development Economics*, *97*(2), 494–504.

Li, H. & Urmanbetova, A. (2007). 14 the effect of education and wage determination in china's rural industry. In *Private Enterprises and China's Economic Development*, (pp. 235). Emerald Group Publishing Limited.

Light, A. & Strayer, W. (2004). Who receives the college wage premium assessing the labor market returns to degrees and college transfer patterns. *The Journal of Human Resources*, *39*(3), 746–773.

Lillo, A. (2006). The private returns to tourist human capital: Endogeneity of schooling and returns heterogeneity.

Lillo-Bañuls, A. & Casado-Díaz, J. M. (2010). Rewards to education in the tourism sector: one step ahead. *Tourism Economics*, *16*(1), 11–23.

Ma, X. & Iwasaki, I. (2021). Return to schooling in china: A large meta-analysis. *Education Economics*, *29*(4), 379–410.

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2022). Robust bayesian meta-analysis: Addressing publication bias with model-averaging.

Maluccio, J. A. (1998). Endogeneity of schooling in the wage function: Evidence from the rural philippines.

Mazrekaj, D., De Witte, K., & Vansteenkiste, S. (2019). Labour market consequences of a high school diploma. *Applied Economics*, *51*(21), 2313–2325.

Miller, P., Mulvey, C., & Martin, N. (1995). What do twins studies reveal about the economic returns to education? a comparison of australian and us findings. *The American Economic Review*, *85*(3), 586–599.

Miller, P. W., Mulvey, C., & Martin, N. (2004). A test of the sorting model of education in australia. *Economics of Education Review*, *23*(5), 473–482.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Mishra, V. & Smyth, R. (2012). Returns to schooling in urban china: New evidence using heteroskedasticity restrictions to obtain identification without exclusion restrictions. Technical Report 33, Department of Economics, Monash University, Discussion Paper.

Mishra, V. & Smyth, R. (2014). Returns to education in china's urban labour market: Evidence from matched employer-employee data for shanghai. In *Urban China in the New Era: Market Reforms, Current State, and the Road Forward*, (pp. 169–183). World Scientific.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, *151*(4), 264–269.

Moretti, E. (2004). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics*, *121*, 175–212.

Morgan, S. L. & Morgan, W. R. (1998). Education and earnings in nigeria, 1974-1992. *Research in Social Stratification and Mobility*, *16*, 3–26.

Mphuka, C. & Simumba, J. (2012). Estimating returns to education in zambia. *African Development Review*, *24*(1), 1–16.

Munich, D., Svejnar, J., & Terrell, K. (2005). Returns to human capital under the communist wage grid and during the transition to a market economy. *Review of Economics and Statistics*, *87*(1), 100–123.

Nakamuro, M. & Inui, T. (2012). Estimating the returns to education using a sample of twins–the case of japan.

Ning, M. (2005). *The economic returns to schooling: Evidence from Chinese twins*. PhD thesis, The Chinese University of Hong Kong.

Okuwa, O. B. (2004). Private returns to higher education in nigeria.

Oreopoulos, P. & Petronijevic, U. (2013). Making college worth it: A review of research on the returns to higher education.

Ozawa, S., Laing, S. K., Higgins, C. R., Yemeke, T. T., Park, C. C., Carlson, R., Ko, Y. E., Guterman, L. B., & Omer, S. B. (2022). Educational and economic returns to cognitive ability in low-and middle-income countries: A systematic review. *World Development*, *149*, 105668.

Patrinos, H. A. (2016). Estimating the return to schooling using the mincer equation.

Patrinos, H. A. & Psacharopoulos, G. (2020). Returns to education in developing countries. In *The Economics of education* (pp. 53–64). Elsevier.

Patrinos, H. A., Psacharopoulos, G., & Tansel, A. (2021). Private and social returns to investment in education: the case of turkey with alternative methods. *Applied Economics*, *53*(14), 1638–1658.

Paweenawat, S. W. & Vechbanyongratana, J. (2015). Private returns to stem education in thailand. *Science*, *100*, 150–000.

Peters, A., Dockery, A. M., & Bawa, S. (2022). Course non-completion and multiple qualifications: re-estimating the returns to education in australia. *Australian Journal of Labour Economics*, *25*(1), 55–80.

Pischke, J.-S. & von Wachter, T. (2005). Zero returns to compulsory schooling in germany: Evidence and interpretation. *The Review of Economics and Statistics*, *87*(3), 467–476.

Psacharopoulos, G. (1982). Earnings and education in greece, 1960â€“1977. *European Economic Review*, *17*(3), 333–347.

Psacharopoulos, G. (1994). Returns to investment in education: A global update. *World development*, *22*(9), 1325–1343.

Psacharopoulos, G. & Layard, R. (1979). Human capital and earnings: British evidence and a critique. *Review of Economic Studies*, *46*, 485–503.

Psacharopoulos, G. & Patrinos, H. A. (2004). Returns to investment in education: a further update. *Education economics*, *12*(2), 111–134.

Psacharopoulos, G. & Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics*, *26*(5), 445–458.

Purnastuti, L. (2013). Instrumenting education and returns to schooling in indonesia. *Jurnal Economia*, *9*(2), 166–174.

Purnastuti, L., Losina, R. S., & Joarder, M. A. M. (2015). The returns to education in indonesia: Post reform estimates. *The Journal of Developing Areas*, *49*(2), 183–204.

Qiu, T. (2007). *Private returns to education: earnings, health and well-being*. PhD thesis, University of Bath.

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

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g.. *Journal of applied psychology*, *79*(4), 518.

Ritchie, S. J. & Tucker-Drob, E. M. (2018). How much does education improve intelligence? a meta-analysis. *Psychological science*, *29*(8), 1358–1369.

Rouse, C. E. (1999). Further estimates of the economic return to schooling from a new sample of twins. *Economics of Education Review*, *18*(2), 149–157.

Sackey, H. A. (2008). Private returns to education in ghana: implications for investments in schooling and migration.

Sakellariou, C. & Fang, Z. (2016a). Returns to schooling for urban and migrant workers in. *Applied Economics*, *48*(8), 684–700.

Sakellariou, C. & Fang, Z. (2016b). Returns to schooling for urban and migrant workers in china: a detailed investigation. *Applied Economics*, *48*(8), 684–700.

Salas-Velasco, M. (2006). Private returns to an university education: An instrumental variables approach.

Salehi-Isfahani, D., Tunali, I., & Assaad, R. (2009). A comparative study of returns to education of urban men in egypt, iran, and turkey. *Middle East Development Journal*, *1*(2), 145–187.

Schultz, T. W. (1961). Investment in human capital. *The American economic review*, *51*(1), 1–17.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Sinning, M. (2014). How much is it worth? new estimates of private returns to university education in australia. Technical report, Final Report.

Sinning, M. (2017). Gender differences in costs and returns to higher education. *AND GENDER*, *2017*, 227.

Sohn, K. (2013). Monetary and nonmonetary returns to education in indonesia. *The Developing Economies*, *51*(1), 34–59.

Staiger, D. & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586.

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

Stanley, T. D. (2005). Beyond publication bias. *Journal of economic surveys*, *19*(3), 309–345.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, *70*(1), 103–127.

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

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox.

Stephens Jr, M. & Yang, D. Y. (2014). Compulsory education and the benefits of schooling. *American Economic Review*, *104*(6), 1777–1792.

Taber, C. R. (2001). The rising college premium in the eighties: Return to college or return to unobserved ability? *Review of Economic Studies*, *68*(3), 665–691.

Troske, K. R. (1999). Evidence on the employer size-wage premium from worker-establishment matched data. *The Review of Economics and Statistics*, *81*(1), 15–26.

Umar, H. M., Ismail, R., & AbdulHakim, R. (2014). Regional disparities in private returns to education: Evidence from nigeria. *Journal of Economics and Sustainable Development*, *5*(20), 48–58.

van Aert, R. C. & van Assen, M. A. L. M. (2021). Correcting for publication bias in a meta-analysis with the p-uniform* method.

Van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*(5), 713–729.

van der Hoeven, R. (2013). Oyedolapo chiamaka adeoye.

Van Praag, M., van Witteloostuijn, A., & van der Sluis, J. (2013). The higher returns to formal education for entrepreneurs versus employees. *Small Business Economics*, *40*, 375–396.

Vasudeva Dutta, P. (2006). Returns to education: New evidence for india, 1983–1999. *Education Economics*, *14*(4), 431–451.

Vivatsurakit, T. & Vechbanyongratana, J. (2020). Returns to education among the informally employed in thailand. *Asian-Pacific Economic Literature*, *34*(1), 26–43.

Walker, I. & Zhu, Y. (2008). The college wage premium and the expansion of higher education in the uk. *The Scandinavian Journal of Economics*, *110*(4), 695–709.

Wambugu, A. (2003). *Essays on earnings and human capital in Kenya*. PhD thesis, University of Connecticut.

Warunsiri, S. & McNown, R. (2010). The returns to education in thailand: A pseudo-panel approach. *World Development*, *38*(11), 1616–1625.

Webbink, D. (2004). Returns to university education.

Wincenciak, L. (2020). Evolution of private returns to schooling over the business cycle in a transition economy. *International Journal of Manpower*, *41*(8), 1307–1322.

Wincenciak, L., Grotkowska, G., & Gajderowicz, T. (2022). Returns to education in central and eastern european transition economies: The role of macroeconomic context. *Research in Comparative and International Education*, *17*(4), 655–676.

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

Zhong, H. (2011). Returns to higher education in china: What is the role of college quality? *China Economic Review*, *22*(2), 260–275.

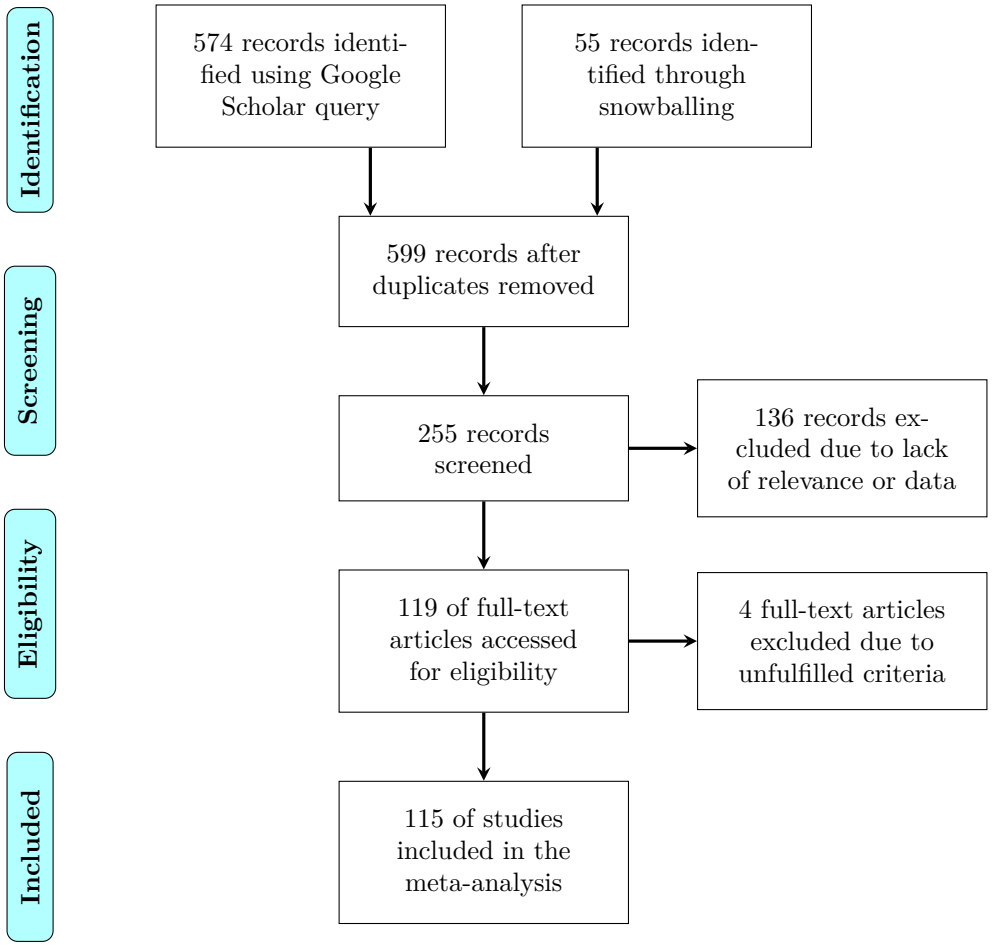
Zhu, R. (2012). Economic restructuring, heterogeneous returns to schooling and the evolution of wage inequality in urban china. In *35th Pacific Trade and Development Conference*.

Ziegel, E. R. (2002). Statistical inference.

Appendix A

Literature Exploration

Figure A.1: PRISMA Flow Diagram



Note: This figure displays a PRISMA flow diagram that graphs the study inclusion process. I use the following Google Scholar query in the search: ***("ability bias" OR "intelligence bias") AND ("private returns") AND ("income" OR "earnings") AND ("schooling" OR "education")***. The query search was conducted during a single day on January 23, 2023. The snowballing was conducted roughly a month later. For the list of the 115 studies included in the analysis, see Table A.1. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. In constructing the diagram, I follow the advice of Moher et al. (2009) and Havránek et al. (2020).

Table A.1: Studies used in the analysis

Panel A: Studies identified by the query	
Acemoglu & Angrist (1999)	Leigh (2008)
Agrawal (2012)	Li & Urmanbetova (2007)
Arkes (2010)	Lillo (2006)
Aromolaran (2006)	Lillo-Bañuls & Casado-Díaz (2010)
Aryal et al. (2022)	Maluccio (1998)
Asadullah (2006)	Mazrekaj et al. (2019)
Aslam (2007)	Mishra & Smyth (2012)
Ayyash et al. (2020)	Mishra & Smyth (2014)
Bakis et al. (2013)	Morgan & Morgan (1998)
Bartolj et al. (2013)	Mphuka & Simumba (2012)
Bergman & Schöön (2018)	Okuwa (2004)
Blundell et al. (2001)	Patrinos et al. (2021)
Botchorishvili (2007)	Paweenawat & Vechbanyongratana (2015)
Campaniello et al. (2016)	Peters et al. (2022)
Campos & Reis (2017)	Purnastuti (2013)
Casado-Díaz & Lillo-Bañuls (2005)	Purnastuti et al. (2015)
Chanis et al. (2021)	Qiu (2007)
De Brauw & Rozelle (2008)	Sackey (2008)
Depken et al. (2019)	Sakellariou & Fang (2016a)
Doan et al. (2020)	Sakellariou & Fang (2016b)
Dumauli (2015)	Salas-Velasco (2006)
Fang et al. (2012)	Salehi-Isfahani et al. (2009)
Fersterer et al. (2008)	Sinning (2014)
Frazer (2023)	Sinning (2017)
Gibson & Fatai (2006)	Sohn (2013)
Giles et al. (2019)	Umar et al. (2014)
Girma & Kedir (2005)	van der Hoeven (2013)
Glewwe (1996)	Van Praag et al. (2013)
Guifu & Hamori (2009)	Vasudeva Dutta (2006)
Harmon et al. (2002)	Vivatsurakit & Vechbanyongratana (2020)
Hawley (2004)	Walker & Zhu (2008)
Himaz & Aturupane (2016)	Wambugu (2003)
Joseph (2020)	Warunsiri & McNown (2010)
Kenayathulla (2013)	Webbink (2004)
Kolstad & Wiig (2015)	Wincenciak (2020)
Krafft (2018)	Zhong (2011)
Krafft et al. (2019)	Zhu (2012)
Panel B: Studies identified by snowballing	
Aakvik et al. (2010)	Heckman et al. (2006)
Angrist (1995)	Hubbard (2011)
Angrist & Krueger (1991)	Ichino & Winter-Ebmer (1999)
Belzil & Hansen (2002)	Ichino & Winter-Ebmer (2004)
Brainerd (1998)	Jones (2001)
Breda (2014)	Kane & Rouse (1993)
Capatina (2014)	Kijima (2006)

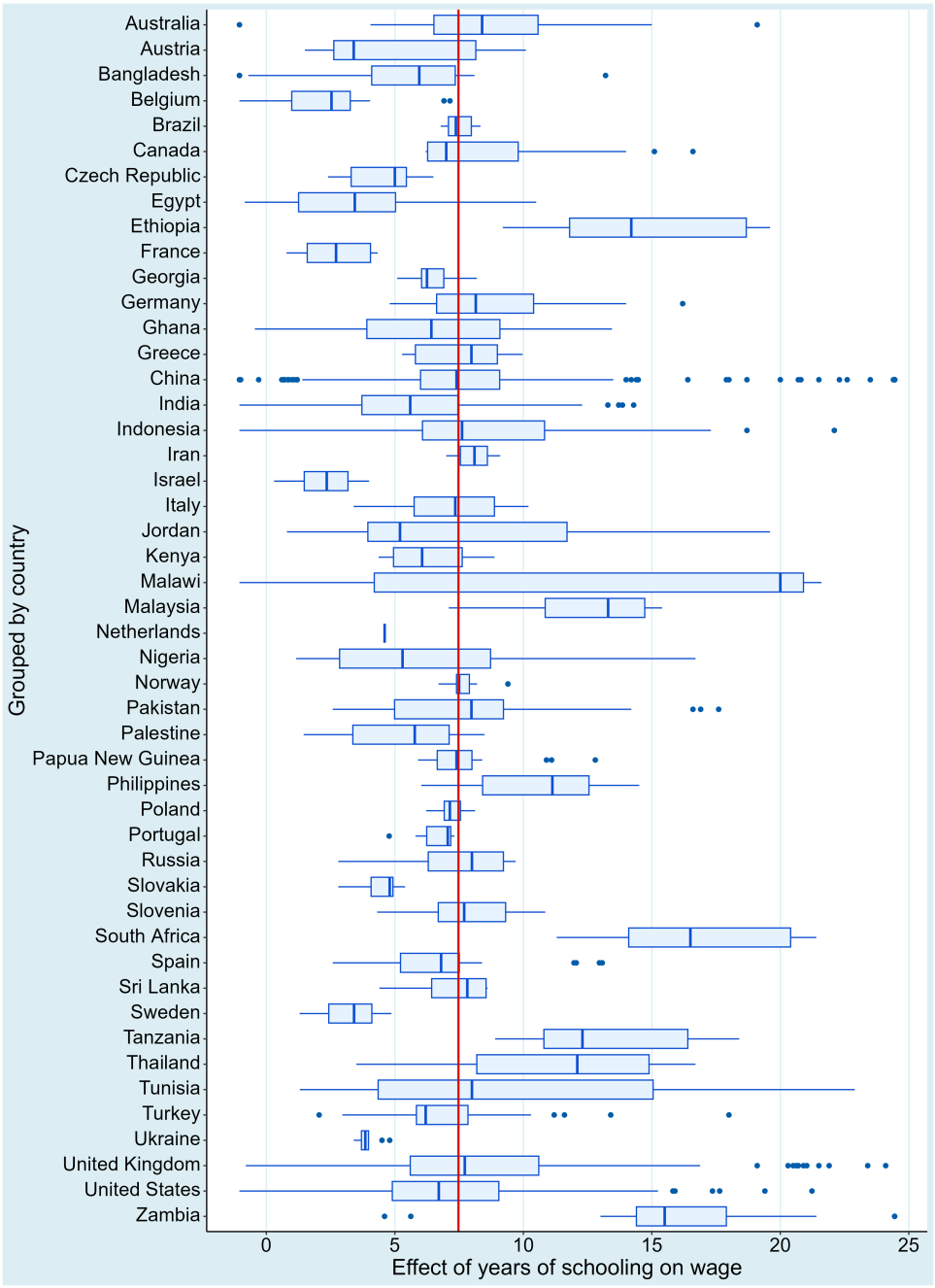
Continued on next page

Table A.1: Studies used in the analysis (continued)

Card (1995)	Kingdon (1998)
Carneiro et al. (2011)	Leigh & Ryan (2008)
Chase (1998)	Lemieux & Card (2001)
Devereux & Hart (2010)	Light & Strayer (2004)
Dougherty & Jimenez (1991)	Moretti (2004)
Duflo (2001)	Munich et al. (2005)
Duraisamy (2002)	Pischke & von Wachter (2005)
Fortin (2008)	Psacharopoulos (1982)
Gill & Leigh (2000)	Psacharopoulos & Layard (1979)
Gorodnichenko & Peter (2005)	Staiger & Stock (1997)
Grogger & Eide (1995)	Stephens Jr & Yang (2014)
Harmon & Walker (1995)	Taber (2001)
Harmon & Walker (1999)	Troske (1999)
Panel C: Twin studies	
Ashenfelter & Krueger (1994)	Isacsson (1999)
Ashenfelter & Rouse (1998)	Isacsson (2004)
Behrman et al. (1994)	Li et al. (2012)
Behrman & Rosenzweig (1999)	Miller et al. (1995)
Bingley et al. (2009)	Miller et al. (2004)
Blanchflower & Elias (1999)	Nakamuro & Inui (2012)
Bonjour et al. (2003)	Ning (2005)
Feigenbaum & Tan (2020)	Rouse (1999)

Note: This table lists all studies used in the analysis. Panel A shows 74 studies identified by the main Google Scholar query; panel B shows 41 studies identified by snowballing. These two panels together present 115 studies from the main dataset, explored in Chapter 3. Panel C displays 16 studies from the twin dataset, explored in Chapter 7.

Figure A.2: Box plot of estimates across countries

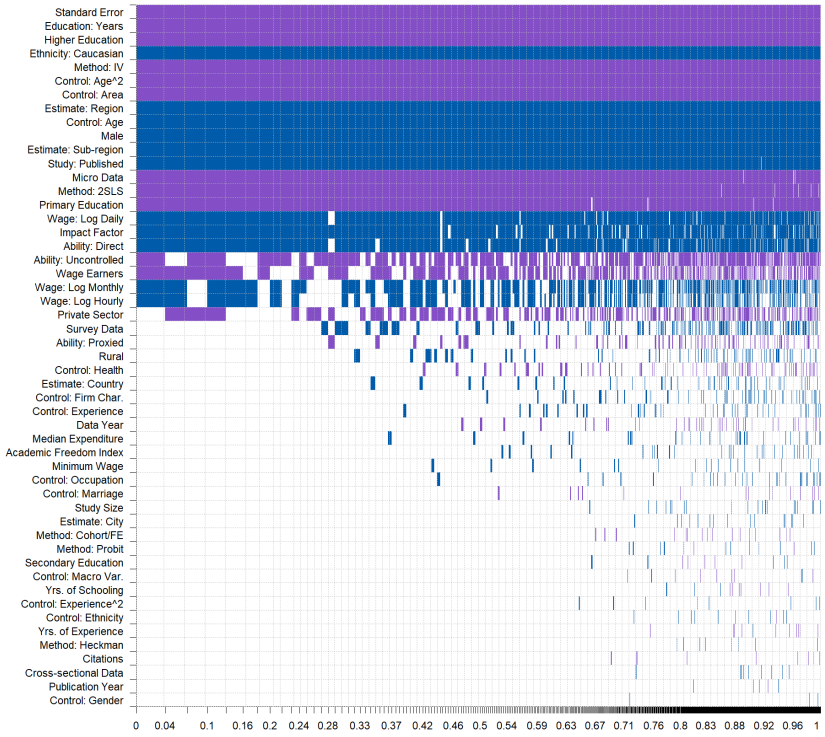


Note: This figure shows a box plot where the reported estimates are grouped at the country level. The data of all 48 countries from the data set is displayed. The red line represents the average effect across the literature. Each box's length represents the interquartile range between the 25th and 75th percentiles. The dividing line within each box indicates the median value. The whiskers extend to the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. Outliers are depicted as blue dots. The data is winsorized at 1% level.

Appendix B

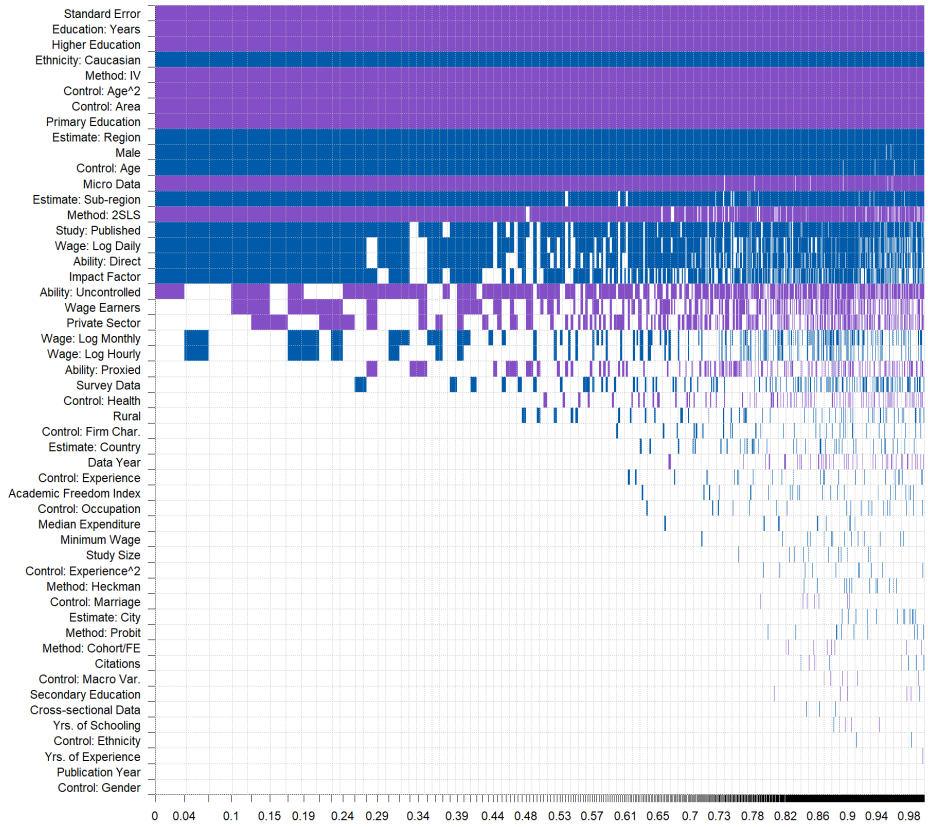
Bayesian model averaging robustness check

Figure B.1: BMA - uniform g-prior and uniform model prior



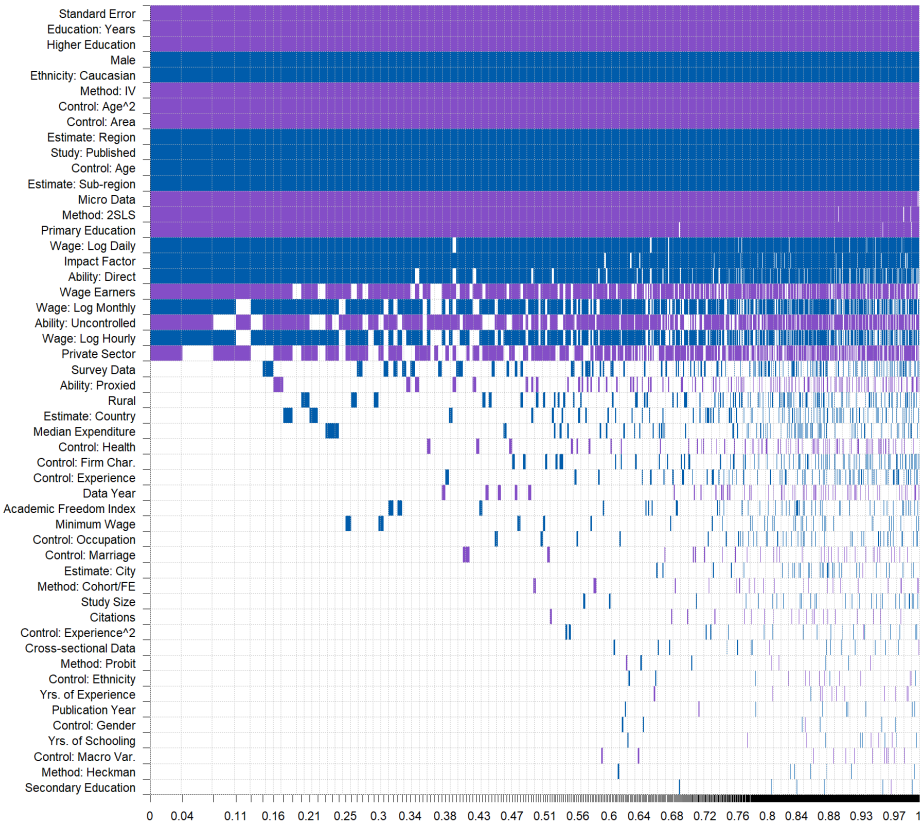
Note: This figure unveils the results of running the Bayesian model averaging using different specifications, namely the uniform g-prior and the uniform model prior. BMA = Bayesian model averaging. For further explanation of the procedure and individual variables, see Figure 5.1 and Table 5.1.

Figure B.2: BMA - benchmark g-prior and random model prior



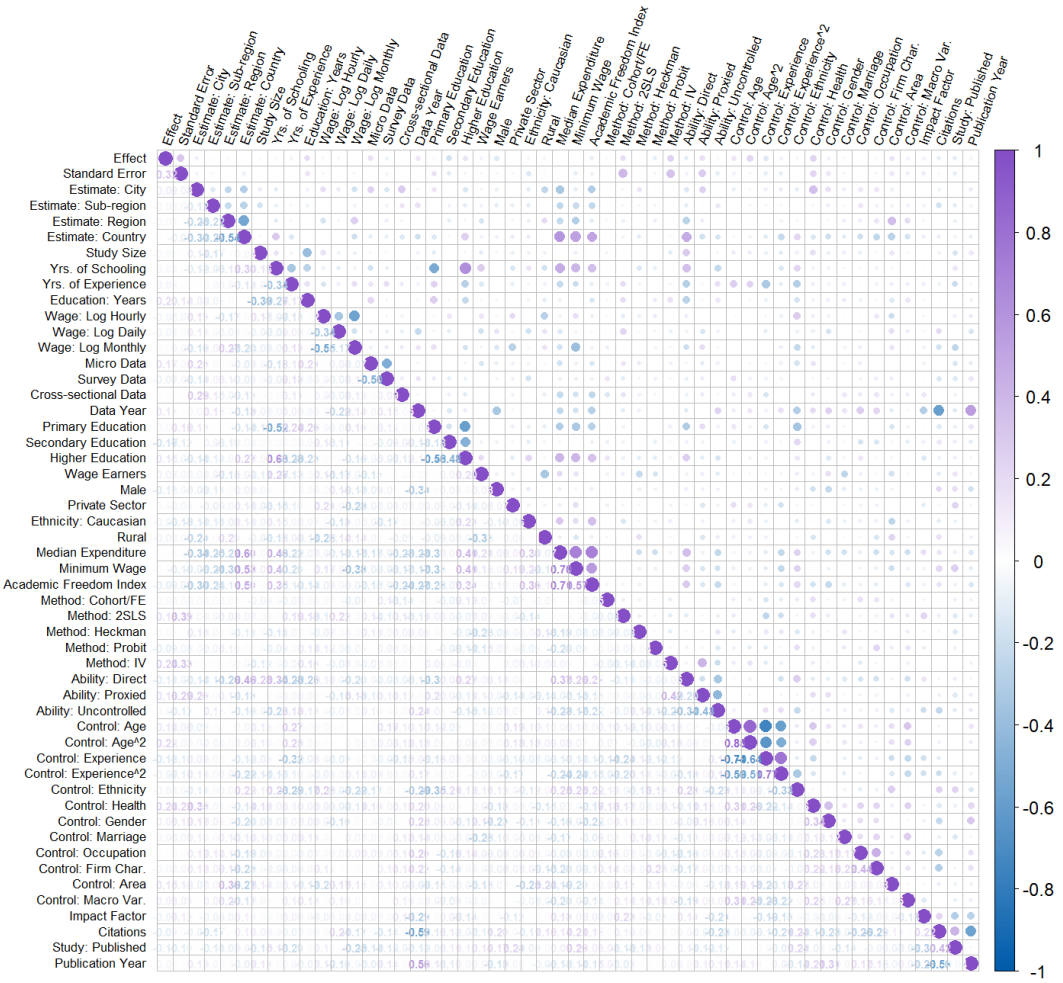
Note: This figure unveils the results of running the Bayesian model averaging using different specifications, namely the benchmark g-prior and the uniform model prior. BMA = Bayesian model averaging. For further explanation of the method and the employed variables, see Figure 5.1 and Table 5.1.

Figure B.3: BMA - HQ g-prior and random model prior



Note: This figure unveils the results of running the Bayesian model averaging using different specifications, namely the Hannan-Quinn criterion g-prior and the uniform model prior. BMA = Bayesian model averaging. HQ = Hannan-Quinn Criterion. For further explanation of the method and the employed variables, see Figure 5.1 and Table 5.1.

Figure B.4: Bayesian Model Averaging - Correlation Table



Note: The figure shows the correlation between variables employed in the Bayesian model averaging. These variables are depicted on both axes. Purple color indicates a positive correlation; blue color indicates a negative correlation. Uniform g-prior and dilution model prior are used in the analysis. For results of the actual estimation, see Chapter 5. For a detailed explanation of the variables used, see Table 5.1.

Appendix C

Implied best-practice across literature

Table C.1: Comparing best-practice estimates across literature

Study	Estimate	95% Confidence Interval
Author's subjective estimate	6.849	(6.098; 7.6)
Panel A: Studies identified by query (subset)		
Leigh (2008)	8.347	(6.838; 9.856)
Bartolj et al. (2013)	8.117	(7.472; 8.762)
Salas-Velasco (2006)	6.381	(5.209; 7.553)
Lillo-Banuls & Casado-Diaz (2010)	7.000	(5.751; 8.249)
Wincenciak (2020)	4.062	(2.839; 5.285)
Okuwa (2004)	6.713	(5.19; 8.236)
Webbink (2004)	10.158	(8.218; 12.098)
Kenayathulla (2013)	9.194	(7.501; 10.887)
Asadullah (2006)	5.781	(4.284; 7.278)
Maluccio (1998)	9.221	(8.041; 10.401)
Depken et al. (2019)	12.190	(10.581; 13.799)
Purnastuti et al. (2015)	10.625	(8.594; 12.656)
Umar et al. (2014)	8.514	(7.095; 9.933)
Sinning (2014)	11.477	(10.43; 12.524)
Agrawal (2012)	7.156	(5.729; 8.583)
Sackey (2008)	6.196	(5.273; 7.119)
Patrinos et al. (2021)	7.593	(6.301; 8.885)
Giles et al. (2019)	6.597	(5.78; 7.414)
van der Hoeven (2013)	7.747	(5.74; 9.754)
Acemoglu & Angrist (1999)	6.546	(5.425; 7.667)
Vivatsurakit & Vechbanyongratana (2020)	10.587	(9.337; 11.837)
Qiu (2007)	5.846	(4.792; 6.9)
Mphuka & Simumba (2012)	12.434	(10.558; 14.31)
Aslam (2007)	10.169	(8.521; 11.817)

Continued on next page

Table C.1: Best-practice across literature (continued)

Study	Estimate	95% Confidence Interval
Himaz & Aturupane (2016)	8.043	(6.336; 9.75)
Warunsiri & McNown (2010)	9.298	(7.583; 11.013)
Aromolaran (2006)	6.633	(5.622; 7.644)
Salehi-Isfahani et al. (2009)	5.032	(4.03; 6.034)
Botchorishvili (2007)	6.285	(5.013; 7.557)
Girma & Kedir (2005)	6.822	(4.727; 8.917)
De Brauw & Rozelle (2008)	7.791	(6.394; 9.188)
Chanis et al. (2021)	8.575	(7.132; 10.018)
Paweenawat & Vechbanyongratana (2015)	10.628	(9.274; 11.982)
Vasudeva Dutta (2006)	3.614	(2.144; 5.084)
Gibson & Fatai (2006)	6.318	(5.166; 7.47)
Hawley (2004)	6.405	(5.074; 7.736)
Sohn (2013)	9.375	(7.932; 10.818)
Harmon et al. (2002)	10.018	(9.024; 11.012)
Lillo (2006)	5.093	(4.137; 6.049)
Zhong (2011)	8.776	(7.135; 10.417)
Krafft (2018)	7.518	(5.57; 9.466)
Walker & Zhu (2008)	9.720	(8.722; 10.718)
Wambugu (2003)	9.906	(7.789; 12.023)
Aryal et al. (2022)	6.996	(5.593; 8.399)
Bakis et al. (2013)	6.087	(5.336; 6.838)
Campaniello et al. (2016)	4.931	(3.604; 6.258)
Joseph (2020)	9.501	(7.706; 11.296)
Dumauli (2015)	10.443	(8.997; 11.889)
Fersterer et al. (2008)	5.062	(3.28; 6.844)
Sinning (2017)	11.524	(10.372; 12.676)
Purnastuti (2013)	4.965	(3.366; 6.564)
Arkes (2010)	7.378	(6.151; 8.605)
Glewwe (1996)	7.346	(5.298; 9.394)
Blundell et al. (2001)	6.033	(4.636; 7.43)
Ayyash et al. (2020)	8.716	(7.138; 10.294)
Panel B: Studies identified by snowballing		
Aakvik et al. (2010)	5.961	(4.683; 7.239)
Angrist (1995)	7.419	(6.417; 8.421)
Angrist et al. (1991)	8.807	(7.625; 9.989)
Belzil et al. (2002)	5.943	(4.828; 7.058)
Brainerd (1998)	2.465	(1.136; 3.794)
Breda (2014)	0.884	(-0.494; 2.262)
Capatina (2014)	6.357	(5.759; 6.955)
Card (1995)	6.127	(5.031; 7.223)
Carneiro et al. (2011)	6.852	(5.251; 8.453)
Chase (1998)	3.115	(2.11; 4.12)
Devereux et al. (2010)	6.210	(4.485; 7.935)
Dougherty et al. (1991)	6.067	(4.677; 7.457)
Duflo (2001)	7.471	(6.215; 8.727)

Continued on next page

Table C.1: Best-practice across literature (continued)

Study	Estimate	95% Confidence Interval
Duraisamy (2002)	6.473	(5.785; 7.161)
Fortin (2008)	4.105	(3.537; 4.673)
Gill et al. (2000)	7.270	(6.486; 8.054)
Gorodnichenko (2005)	4.944	(4.029; 5.859)
Grogger et al. (1995)	3.367	(2.567; 4.167)
Harmon et al. (1995)	10.136	(8.619; 11.653)
Harmon et al. (1999)	9.261	(7.664; 10.858)
Harmon et al. (2003)	7.878	(6.688; 9.068)
Heckman et al. (2006)	8.440	(7.372; 9.508)
Hubbard (2011)	7.005	(6.174; 7.836)
Ichino (1999)	7.507	(6.123; 8.891)
Ichino et al. (2004)	10.498	(8.981; 12.015)
Jones (2001)	6.154	(4.512; 7.796)
Kane et al. (1993)	6.584	(5.739; 7.429)
Kijima (2006)	2.899	(2.039; 3.759)
Kingdon (1998)	7.976	(6.553; 9.399)
Leigh (2008)	8.277	(7.132; 9.422)
Lemieux et al. (2001)	6.230	(5.321; 7.139)
Light et al. (2004)	8.067	(7.001; 9.133)
Moretti (2004)	6.581	(5.074; 8.088)
Munich et al. (2005)	5.043	(3.653; 6.433)
Pischke (2005)	6.801	(5.282; 8.32)
Psacharopoulos (1982)	3.714	(2.318; 5.11)
Psacharopoulos (1979)	7.461	(6.058; 8.864)
Staiger et al. (1997)	7.507	(6.298; 8.716)
Stephens Jr et al. (2014)	6.126	(4.95; 7.302)
Taber (2001)	6.277	(5.119; 7.435)
Troske (1999)	3.728	(2.574; 4.882)

Note: The table reports estimates of the implied best-practice across studies of the main dataset, as well as the author's subjective best-practice. For clarity of presentation, I arbitrarily removed several query-identified studies from the table. 95% confidence interval bounds are constructed as an approximate using OLS with study level clustered standard errors.