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 down the line in the form of increased earnings stemming mainly from increased productivity of the individual (Schultz 1961). Mincer (1974) then 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 quadratic experience term. Rigorously, we can write this semi-logarithmic relationship as: ln(Yi) = α + βSi + γ1Xi + γ2X 2 i + ϵi , (2.1) where ln(Yi) denotes the log of earnings of an individual i, Si represents their attained years of schooling, Xi stands for the years of work experience of said individual, and lastly, ϵi captures the individual-specific error. In case that individual’s experience is absent, Mincer (1974) proposes using a measure of potential experience instead. This can be calculated as: 2. Private returns to education 3 Xi = Ai − Si − 6, (2.2) where Xi is the potential experience, Ai represents the individual’s age, and Si stands for their finished years of schooling. 6 is used as a constant with the assumption that the individual starts their education at age 6. Over the decades, this equation has been subjected to scrutiny of heaps of research papers (Ashenfelter & Krueger 1994; Heckman et al. 2006; Card & Krueger 1992) and naturally, this scrutiny raised a number of 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 arrives 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. This highlighted the importance of an individual’s location factors and the role they play 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 strongly back up their claim. Belzil & Hansen (2004) then extend the equation to account for individual heterogeneity by using a dynamic programming model of schooling decisions. Among a plethora of methods for estimating this relationship, OLS stands as the most common approach (Ashenfelter et al. 1999; Card 1999). However, OLS estimates suffer from several estimation problems including sample selectivity, the omitted variable bias, or the measurement error bias, as noted for example by Aslam (2007). 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 not to tackle, and that I 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 their lack of could be causing biased estimates - individual ability. There exists a plethora of research in psychology to show that general 2. Private returns to education 4 intelligence is one of the most reliable predictors of one’s success (Gottfredson 1997; Deary et al. 2007; Deary 2020; Ozawa et al. 2022). If we quantify this measure (so called g-factor), 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, and find out that they serve as crucial predictors in these areas of economic development. When it comes to policy making, the predictive analysis prevalent in psychology is not as helpful (Almlund et al. 2011). While highly useful when placing an individual into the labor market, the 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 changes in policy, it is impossible to evaluate the quality of said change. Without a doubt, 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 easily 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. The term ability bias describes this phenomenon and has over the years 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; rather it may introduce new biases on its own. A whole new branch of research into ability bias lies too within natural experiments. 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 2. Private returns to education 5 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 certain threat to the result validity, as those errors could end up explaining the majority of twin-level differences across the population (Ashenfelter et al. 1999). Nonetheless, twin studies provide an interesting alternative way to survey the ability bias issue from yet another perspective, although most authors choose to overlook this possibility completely. On balance, the ability bias issue remains in a somewhat niche spot of researchers’ consciousness. Given the lack of consensus in 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). Number of times, the authors acknowledge that their estimates could be plagued by the bias, but fail to obtain the data necessary for its treatment (Agrawal 2012; De Brauw & Rozelle 2008). Other times the issue gets overlooked completely and the authors simply focus on either the simplest or a more complex form of the Mincer regression (Angrist 1995; Sinning 2017). Given the disunified practice, I find it of high importance 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 trying to answer the questions, I find it crucial to look at and acknowledge the existing meta-analyses that might 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 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 10 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 when controlled for, ability turns out to be a significant predictor of returns to schooling (about 0.8-0.9% points). They conclude that omission of ability bias may indeed 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 analyse the issue in depth. 5 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 latter two do not. Lastly, 6 out of the 10 existing meta-analyses include in any form a control for methodology in their approach. Mostly 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 within each other. Indeed, there is not a single study that would bring all these issues together and try to answer all of them. This, among with other important points, should be the main focus of this thesis, as explained in the following section. 2. Private returns to education 7 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 (this paper has actually been published only after the conceptualization of this thesis). Although I may not be the first to consider ability bias as an important predictor to 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 explore the way 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 will now be able to 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 perhaps highlight the importance of obtaining data through which the researchers can control for this bias. Given the existing heterogeneity in the existing 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 methods of estimation may systematically have on returns to education. Despite the fact that over a half of existing meta-analyses address this issue, none of them provides a direct comparison of all the existing methodology within the literature. Given that the dataset I will assemble and use to test for this relies primarily on a search query for choice of studies, the literature set should provide the most representative form of the existing literature possible and capture most of the important 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 stateof-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 attempted to answer the issue, no consensus exists here either 2. Private returns to education 8 (three claim the presence of publication bias, while two argue for the lack of thereof). More scrutiny should help clear out the uncertainty about publication bias and provide even more robustness to the results. Fifth, I will take a look at the role of individual variables in regards to the effect behavior with the use of novel technology including Bayesian and Frequentist Model Averaging. Which variables are most important drivers of the returns to education effect? What is their economic significance? If a bestpractice effect could be derived from the literature that would correct for the aforementioned detected biases, what would be the true effect? None of the existing research tackles any of these questions, and I hope this approach will contribute to their clarification. Sixth, I should construct an additional, entirely new dataset including only natural experiments conducted on twins (so called twin studies) and run the analysis again 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 the role that education plays in affecting the twins’ future earnings. Furthermore, I can observe how all the answers to the aforementioned questions regarding ability and publication bias, heterogeneity, and so on, change when using this new twin study dataset. This should help me validate the robustness of my results, and identify the most important ones. Next, I present several technical extensions as an improvement to 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 only exists in the STATA software. I hope to facilitate research to a surely sizable pool of researchers who do not work with or hold the license to STATA by providing the code for the aforementioned method purely in the programming language R. With numerous validity checks to ensure the quality of the code, I believe this small task may aid further researchers to effortlessly validate the robustness of their results. As the second technical extension, I upgrade the existing code of the STEM method (Furukawa 2019) to work in orders of 35 times faster than the available source code.1 As the last extension, I provide an all-encompassing R code in form of 1Tested on the 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 10 runs. 2. Private returns to education 9 several scripts that can be used together to easily replicate the whole analysis.2 With over 5000 lines of code, the project allows the user to run, see, and customize every single method from within a single script. All results then get automatically exported and saved in a single, small-sized, 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 the complex meta-analysis methods and run them all in a matter of seconds. In fact, tested against code from my bachelor thesis that took 9 minutes to run, the newly presented code is able to run in 2.31 seconds (16.77 uncached). This corresponds to an improvement factor of over 230x for the cached version (30x for the uncached). To assemble my dataset, I employ the Google Scholar search engine due to its full-text search capabilities. A query constructed using a combination keywords helps me narrow down the results into studies dealing with ability bias, private returns, and education. After several modifications in order to ensure that the query generates consistent results within the scope of interest, I obtained the final form which can be found in 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 to note down key information about all 574 of these studies. These included the authors’ names, publication information, the number of citations and the impact factor of the journal the study was published in1 . I also extracted the result ID of each study in order to guarantee the uniqueness of each hit, and to avoid duplicate results. I then go through the first 200 of the studies, and identify 122 of them eligible for data collection. This means they report an estimate of a regression of wage on a schooling variable. This preliminary check suggests that over 60% of the studies returned by are query are relevant to the topic, validating 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 study list into its final form. For a study to be included in the dataset, it needs to fulfill several criteria. 1 In case of an unpublished study, I set the impact factor to 0. 3. Data 11 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 obvious that the sample size corresponds to the reported estimates. 48 studies do not fulfill these criteria and are thus removed which leaves 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 clear consensus in existing literature on which approach should be taken, as highlighted by Stanley (2001). Although they advise careful consideration when including unpublished studies, they also acknowledge that their omission could create publication bias on its own. The inclusion approach is also supported by Cook et al. (1993), who found that numerous meta-analysis researchers and methodologists believe that data from unpublished studies should not be discarded if one wishes to objectively synthesize the available information. 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 clearly encompass the whole field of relevant literature, I employed 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 a number of 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 studies (Card 1995; Heckman et al. 2006; Psacharopoulos & Patrinos 2018) have presented results that have since been many a times reviewed and are indeed well-established, the omission of some of these crucial studies seems 3. Data 12 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, and discovered 55 additional papers that greatly contribute to the topic of returns to education. After applying the aforementioned criteria, I was able to further narrow down this list to 41 highly relevant and collectible papers. When combined with the 74 studies identified during the query search, the final list consist 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 overall provide a more robust representation than the query search subset. The final list of studies can be found in Appendix A. 3.2 Interpreting of the Effect in Question A quick glance into the assembled literature set reveals an important issue that 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 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(Yi) = α + β1P RIMi + β2SECi + β3HIGHERi + γ1Xi + γ2X 2 i + ϵi , (3.1) where P RIM, 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 Bachelor’s degree, Master’s degree, or even country-specific education levels. The important question here is - having these different levels of schooling, 3. Data 13 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, we can follow the example of (XXX) and quantify the relationship between schooling in levels and years of schooling in the following formula: Si = (1 + βi,higher − βi,lower) 1 Yi,higher−Yi,lower − 1, (3.2) where Si denotes the effect an additional year of schooling has on the log wage of an estimate i, βi,higher and βi,lower are the coefficients from the Mincer regression associated with the higher and lower schooling levels respectively, and finally Yi,higher and Yi,lower are the number of 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). In case no other level is available for comparison (such as when calculating the returns to schooling for the first level coefficient in the equation), we can simply plug 0s for the other schooling level. This reduces the equation to the following form: Si = (1 + βi) 1 Yi − 1, (3.3) where βi is the Mincer regression coefficient associated with the attained schooling level of an estimate i, and Yi denotes the number of years required to attain said level. After transformation of 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. After obtaining both the transformed estimates and their standard errors, I only then calculate the t-statistics, which ensures the validity of publication bias methods used later in the work. In order to allow easier presentability, I scale all the numbers by a factor of 100 so that the effect 3. Data 14 can be directly interpreted as a percentage return to an additional return of schooling. The last important question to answer is whether, after unifying the different types of 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 years of schooling vary systematically depending on the type of education measure used. They find that studies using years of schooling report higher estimates than those using education levels. Given this finding, I choose to to include a variable in my dataset that controls for the type of estimate reporting used. In theory, such coefficient should be 0 (meaning there is no systematic difference between reporting in years and in levels), and in Chapter 5 we will be able to see whether this holds true. 3.3 Dataset assembly With the effect interpretation clearly outlined, I got to the actual data collection. From the 115 relevant studies, I collected a total of 1754 estimates of the effect together with dozens of other variables that help me capture heterogeneity within the literature. Apart from the necessary numeric statistics such as standard error, t-statistic, or the degrees of freedom, I also collect over 40 variables categorizing the effect type, study characteristics, spatial/structural variation, estimation method, and publication characteristics. The list of all these variables, along with their descriptions and summary statistics, can be found in Table 5.1. Upon closer inspection, I observed that the studies in my dataset can be clearly categorized into 4 categories based on their approach to the issue of ability bias. 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. Oftentimes, this is 3. Data 15 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 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 able to successfully collect the data for a vast majority of them, although some variable groups still had to be dropped for the lack of thereof. Topics such as education field (STEM, Medicine, Law,...) regression type (Mincer vs. Discounting), or school type (Private vs. Public), were all address within but a few if any studies, making them infeasible to collect. On the flip side, I was able to identify and add in a handful of variables that I had not intended to add initially, such as marriage control, or residential area type (rural vs. urban). I also added data on 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). When it comes to study specific variables including the number of citations, publication status, or impact factor, I made sure that all these could be directly comparable by measuring them in a single day - January 23th, 2023. Any changes within these variables for the included studies after this date are not considered. I can further make use of the human capital earnings function described in of 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 mean years of schooling, mean years of experience, or mean age, provided the other two are reported. For example, if a study reports mean age and mean years of schooling of the subjects, but omits the mean experience, it can be calculated as age - schooling - 6. On the flip side, there are times when a 3. Data 16 study fails to report at least two of these variables. In those cases, I leave the underivable values empty. However, later methods used in this work require there to be no missing observations in the data whatsoever. To treat this, I make use of clever interpolation, and 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 of dummy type, such as wage earners vs. self-employed, I use the median of the existing data for the given variable. In case the variable can be aggregated on country level, the interpolation happens at that level as well, 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), mean of the data is used, aggregated again on country specific level. This ensures that the ratios always sum up to 1, but at the same time capture the situation representing the environment in which the study took place 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 3.4 Initial analysis After cleaning the dataset and double checking that all calculations are 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. As a baseline for the rest of the work, we can observe that the unweighted mean of the effect across all data is equal to 7.476 (7.652 for the data weighted by the number of estimates reported per study). As a reminder, we can interpret this as a 7.456% increase in log wage per additional year of attained schooling. This falls well into the expected range when compared to other works and serves 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 findings also align well with the only study dealing in detail with ability bias, Wincenciak et al. (2022), who also report a 7% figure for the average effect. Note that the suggested figure of 7.476 does not in any way account for publication bias and should thus be treated only as a benchmark for further comparison. When it comes to other subsets of data, it is impossible to make any concrete conclusions due to the large confidence interval. Despite this, there seems to be a bit of variety in several interesting categories including the age of data, economic status of countries, study publication status, or, perhaps more interestingly, ability. Still, this differences are first of all only marginal and can not serve as any concrete evidence to a clear trend just yet, and second, mean is not the most telling of statistics and I will need to subject the data to more scrutiny in Chapter 5. To highlight the differences between individual studies, I also include box plot of study-level clustered data in figures 3.1 and 3.1 (in the Appendix B, 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), mostly the studies report results close to the mean and no single box appears immediately suspicious. With these box plots, I made sure to double check the results to spot any potential miscalculations and after this validation deemed the dataset as final. In this chapter, I will attempt to uncover any potential publication bias in the primary studies, as per the meta-analysis guidelines by Havránek et al. (2020). The fact that Countless researchers The first way I test for the presence of publication bias is by means of 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) also argues that alternatives can be used, such as square root of the degrees of freedom. 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. Obvious gaps or holes in this funnel may then hint that a data tampering of some kind exists within the literature. As mentioned above, I construct the funnel plot using the inverted standard error as the measure for precision due to the fact that all estimates in the dataset do 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 add in an additional plot that displays only medians of the effect for all 74 studies. These two graphs can be found in the sub-figures of Figure 4.1. No obvious asymmetry nor holes appear at a first glance as far as the plot with all observations is concerned. 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 important takeaway from the latter plot is the lack of suspicious outliers in either directions. 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 rather normal. 4.1 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 in a quantitative and more robust way, I turn first to the Funnel Asymmetry Tests (FAT) (Egger et al. 1997; Stanley 2005; 2008). These test for the funnel plot asymmetry be means of a simple equation where the effect is regressed on the standard error in order 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, which essentially indicates the presence of publication bias. Algebraically, the relationship can be written as: 4. Publication bias 23 Sij = β0 + β1 ∗ (SES)ij + uij , (4.1) where S represents the returns to schooling effect for the i-th observation of the j-th study in the dataset, and SES corresponds to the effect’s standard error. The slope coefficient, beta1, then measures the amount of publication bias in the data, while the intercept coefficient, β0, captures the "true" effect of returns to schooling corrected for publication bias. uij stands for the error regression term. In the tables that follow, 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 weight 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 6.741 percent. This indicates that the underlying effect might be slightly lower than what the simple average of estimates would suggest, approximately by one percentage point. Furthermore, the lowest predicted value can be associated with the study-size weighted model (6.294), suggesting perhaps that larger studies drive the effect upwards. However, it is important to note that this difference is relatively small compared to the other estimates. The discrepancy between the study-size weighted model and Fixed effects model, the latter of which predicts the highest returns to education at 6.741 percent, is less than half a percentage point. 4.2 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 important to acknowledge that this assumption does not always hold. In cases where the relationship behaves in a less straightforward manner, the FAT 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 6 techniques that relax the linearity assumption. The first of these 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 studies that are adequately powered. "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 more reliable it can be considered as. In technical terms, the power is calculated using statistical significance of estimates and then compared to their standard errors. As per 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, staistically speaking, rather paradoxical approach stems from the idea the most researchers use statistical significance as the main benchmark for deciding whether to publish the estimate or not. Stanley et al. (2010) show that if most of the less precise estimates are discarded, the publication bias within the data sample drops greatly. In my data, 10% of estimates corresponds 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, 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%, which is 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 study level. The model parameters themselves get treated as random variables instead of fixed numbers which allows 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 modelling. This complex framework of multi-level modelling 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). 4. Publication bias 26 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 threshold of statistical significance is crossed, the publication probability changes. Andrews & Kasy (2019) then demonstrate how this probability can be calculated in a non-parametric manner, and utilizing the inverse of this probability as new weights, they obtain a non-biased distribution of the estimates. Utilizing 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 have a considerably higher likelihood of being 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 for sufficiently large studies the publication bias is usually absent, the EK model finds a cutoff value below which publication bias should not appear. Bom & Rachinger (2019) then fit a piecewise linear regression with a kink at said cutoff point which allows 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 (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 range of 6 to 7 percent. Only the stem-based method suggests a coefficient for returns to education above 7% (specifically, 7.2%). This is in line with the linear approach and confirms the behavior observed thus far. The Hierarchical Bayes indicates a strong presence of publication bias, while the Endogenous Kink method result is an insignificant one. Finally, the Selection model proposes that results at the 5% significance level have a considerably higher chance for publication than insignificant ones. Up until now, the publication bias tests have been based on the assumption that the correlation between the effect and the standard error is itself an indicator of publication bias. However, this introduces, by definition, endogeneity into the equation. To see how this issue can be treated, it is important to understand how it arises in the first place. The correlation (and thus endogeneity) in the data can come from several sources. First, it could a simple measurement error or wrong calculation procedure that introduce correlation into the data, as the standard error in itself is an estimate. 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 with the goal of improving significance. And lastly, any unobserved heterogeneity may also introduce correlation, this time in the form of inherent differences in methodology that may pose a systematic influence on 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). For an IV regression, we first need an instrument. The criteria for finding a valid one are rather simple and straightforward - 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 (β1 from Equation 4.1) not poisoned by endogeneity. Several instruments appear as valid choices from the IV re- 4. Publication bias 28 gression. These include 1 √nobs , 1 nobs , 1 n 2 obs , and log(nobs), where nobs stands for the number of observations associated with each estimate. There are several inherent properties of the number of observations that makes these instruments valid options. Firstly, the size of an experiment, or the number of subjects in the study, does not directly change the actual effect that exists in the population. If such true effect exists, it should be independent of how many subjects we include in our study. 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 what the results would be had the whole population been used for calculation. Still, out of these four proposed instrument, which one is the best? To find out, I write a helper function in R that automatically proposes a best performing instruments based on the results of several specification tests. These are, namely, the Underidentification test, the Weak identification test, the StockYogo 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 log(nobs) performed the best out of the four instruments. The IV regression, using this instrument, 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 with no prior assumption 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 pvalues 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 puniform\* 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. While the instrumental variable approach proposes rather sensible results, the p-uniform\* is an outlier among previous estimates. This is perhaps more perplexing given the fact that, were between-study variance to be the cause of the effect overinflation, p-uniform\* is precisely the type of method that should account for this. Possibly, these results may be a consequence of either some type of calculation error, or perhaps a hidden trend or anomaly within the data which is hard to detect. Suffice to say, I dug into the calculation multiple times to validate that all specifications and other inputs are sensible, yet was unable to detect anything out of the ordinary. As such, I present the results with a grain of salt, but believe them to be fully valid. 4.4 Caliper tests The next method I will make use of to search for deviations from normality in the reported literature results are Caliper tests, developed by Gerber et al. (2008). The method they propose also does not assume any prior relationship between the effect and the standard effect, similar to the tests from Section 4.3. Here, t-statistics are subjected to scrutiny, and the authors argue, that upon looking at the immediate vicinity of an important significance level, there should occur no structural breaks in the distribution of t-statistic. In other words, the t-statistic distribution should be, in theory, behave rather 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 that appear around significant t-statistic thresholds, such as 1.96, or 1.69, in intervals of varying widths, called Caliper widths. If, within any half of that interval, there is a significant imbalance in the number of tstatistics 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, 4. Publication bias 30 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 along with their standard errors, and hence, t-statistics. Unfortunately, this practice inadvertently promotes emphasis on 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 tampering 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 as described in the previous paragraphs. Two quick notes about the results are in order. First, there are very few (in fact 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 a practical perspective, however, this makes a lot of sense if we presume that the true effect indeed lies around 7%. Given the consistency of the tests carried out in the previous sections, this assumption appears quite feasible. 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, when describing the 1%, 5%, and 10% thresholds respectively. Speaking more bluntly, the words high, medium, and low could be used. It appears that 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 rather sceptical views on the presence of publication bias in the dataset, but perhaps these thresholds could represent initial tangible indications of reporting misbehavior. 4.5 Novel tests for detecting publication bias As a last chapter of my hunt for publication bias, I present 3 new methods that explore the issue of publication bias further. The first two of these methods deal with p-hacking and have both 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 deals with analysing 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 4. Publication bias 32 Model Averaging (RoBMA) (Maier et al. 2022), a technique that is able to 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 then 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 not nearly enough 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 highly important 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 metric has to be calculated by the author using reported standard errors, and can thus be easily p-hacked. Second, even small amounts of p-hacking can have a large impact on the results. Precision is often used as a weighting metric in methods such as the 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 the 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 provides a summary of the individual components - effect, heterogeneity, and publication bias. The effect estimates propose a rather 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 4 models used to estimate individual components2 , 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 there is a positive effect of returns to schooling (5.736-9.520), and most of them 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 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 there exist sizeable jumps 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.