CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

How Does Peer Socioeconomic Status Affect Academic Achievement? A Meta-Analysis

Bachelor's thesis

Author: Tereza Markalousová
Study program: Economics and Finance
Supervisor: Prof. PhDr. Tomáš Havránek Ph.D.
Year of defense: 2023

Declaration of Authorship

I hereby declare that I compiled this thesis independently, using only the listed resources and literature and that the thesis has not been used to obtain any other academic title. I declare that Generative AI tools were used to enhance the writing style of this thesis. The results generated by AI were used in accordance with principles of academic integrity.

I grant to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agree with the thesis being used for study and scientific purposes.

Prague, July 17, 2023

Tereza Markalousová

Abstract

Using state-of-the-art meta-analysis methods, we investigate the effect of peer socioeconomic status (SES) on academic achievement. Our data set covers almost 40 years of research, containing 449 estimates from 40 studies. We examine publication bias for the first time in this research area. Various statistical tests reveal the presence of publication selection and suggest a smaller effect size in comparison to prior findings. Employing Bayesian and frequentist model averaging techniques, we identify factors that systematically influence the magnitude of the estimated effect. Publication bias exerts the strongest upward effect, along with the use of parental education and home resources as measures of SES and combined measures of SES. Conversely, variables such as the number of citations, publication status, science test type, and the use of advanced methods exhibit a negative effect.

Keywords Socieconomic status, Meta-analysis, Social class, Academic achievement

Title How Does Peer Socioeconomic Status Affect Academic Achievement? A Meta-Analysis

Abstrakt

Za pomoci nejmodernějších metod metaanalýzy zkoumáme vliv socioekonomického statusu (SES) vrstevníků na akademické výsledky jednotlivce. Náš dataset zahrnuje téměř 40 let výzkumu a obsahuje 449 odhadů ze 40 studií. Poprvé v této oblasti výzkumu zkoumáme publikační zkreslení. Řada statistických testů indikuje přítomnost publikační selekce a naznačuje menší velikost efektu ve srovnání s předchozími zjištěními. Za pomoci bayesovského a frekventistického průměrování modelů identifikujeme faktory, které systematicky ovlivňují velikost efektu. Nejsilnější pozitivní vliv má publikační zkreslení spolu s použitím vzdělání rodičů a domácích zdrojů jako ukazatelů SES a kombinovaných ukazatelů SES. Naopak proměnné jako počet citací, publikační status, použití testu z vědy a využití pokročilých metod vykazují negativní efekt.

Klíčová slova Socioekonomický status, Metaanalýza, Sociální vrstva, Akademické výsledky

Název práce Jak socioekonomický status vrstevníků ovlivňuje akademické výsledky? Metaanalýza

Acknowledgments

I am especially grateful to Prof. PhDr. Tomáš Havránek Ph.D., who introduced me to a new field of research and provided a variety of interesting methods. Furthermore, he always responded in a timely manner and brought numerous helpful recommendations. I am also tremendously thankful to Petr Čala for his generous contributions of code and insightful advice. Lastly, I would like to acknowledge my family and partner for their unwavering support throughout the research process.

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

Markalousová, Tereza: *How Does Peer Socioeconomic Status Affect Academic Achievement? A Meta-Analysis*. Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2023, pages 76. Advisor: Prof. PhDr. Tomáš Havránek Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AA Academic Achievement
BE Between Effects
BMA Bayesian Model Averaging
BPE Best Practice Estimate
EK Endogenous Kink
FAT Funnel Assymetry Test
FE Fixed-Effect
FLE Free Lunch Eligibility
FMA Frequentist Model Averaging
GAA General Academic Achievement
HLM Hierarchical Linear Modelling
IV Instrumental Variable
MAIVE Meta-Analysis Instrumental Variable Estimator
OECD Organisation for Economic Co-operation and Development
OLS Ordinary Least Squares
OVB Omitted Variable Bias
PISA Programme for International Student Assessment
PIP Posterior Inclusion Probability
PET Precision Effect Test
RE Random Effects
SE Standard Error
SES Socioeconomic Status
VIF Variance Inflation Factor
WAAP Weighted Average of Adequately Powered
WLS Weighted Least Squares

Chapter 1

Introduction

Since the influential report by Coleman (1966), the relationship between academic achievement and socioeconomic status (SES) has been studied in a variety of academic disciplines. While previous research has established a positive link between student's SES and academic performance (Sirin 2005), studies examining the impact of peer influences have produced inconsistent results.

The study of peer effects is based on the notion that, apart from teachers, children learn and develop attitudes from their peers and the school environment. Schneeweis & Winter-Ebmer (2007) emphasized the significant role of peer groups as a source of inspiration and motivation. Within this context, the peer effect of socioeconomic status suggests that a student's academic success can be positively influenced by attending school with peers from higher SES backgrounds. Conversely, it may be academically harmful to attend school with students from lower socioeconomic backgrounds.

If the effect size is substantial, it would have implications for the optimal design of educational systems. Research demonstrates that changes in school organization can significantly impact academic achievement (Betts 1998; Wößmann 2003). Therefore, recognizing positive peer effect of socioeconomic status provides a strong incentive for implementing policy measures to address the resulting achievement gap (Paloyo 2020). Without such interventions, the educational system may fail to fully harness the potential of low-SES students, leading to inefficiencies and further perpetuating societal inequality. Academically, peer effects are essential for the school choice and ability grouping debates, both of which are discussed later in the thesis. In practice, the study of peer effects also yields insights for school principals and parents.

The existing literature on the subject has presented inconclusive evidence, with studies reporting results ranging from a significant effect (Robertson & Symons 2003) to no effect at all (Evans *et al.* 1992). The lack of consensus in the literature motivates the need for a meta-analysis to determine the magnitude of the true effect and identify the factors contributing to the effect size variation.

In simple terms, a meta-analysis is a statistical method that combines data from multiple individual studies to examine underlying relationships and potential biases (Feldman 1984). This thesis expands on the meta-analysis conducted by Van Ewijk & Sleegers (2010) and provides a more comprehensive analysis of the effect of peer socioeconomic status. We employ state-of-the-art methods to address publication bias and model uncertainty. We then provide a robustness check, examine the factors influencing the effect size, and present our subjective best-practice estimate. Furthermore, our paper adheres to guidelines for meta-analysis in economics by Havránek *et al.* (2020).

Our contribution stems from the creation and analysis of a unique data set comprising 449 estimates from 40 studies. Additionally, our study rigorously examines publication bias in this research area for the first time. The application of 17 distinct methods mostly reveals the presence of publication bias, and after accounting for bias, the effect size ranges from 0.117 to 0.338. Through model averaging, we identify several key variables that drive the effect, including standard error, operationalization of the SES variable, academic achievement test type, number of citations, and publication status. Our findings of publication bias contrast with those of Van Ewijk & Sleegers (2010), who found little to no publication bias. This discrepancy in results is also reflected in the differences between our proposed best-practice estimates (BPE) of the effect. While the previous study suggested an estimate of 0.31, our analysis leads us to propose a BPE of 0.2. On the other hand, our examination of the factors influencing the effect size aligns to some extent with previous research, with the main distinction being the inclusion of a broader range of variables in our analysis.

The remaining sections of the thesis are organized as follows: Chapter 2 provides a brief overview of the topic and the literature. Chapter 3 describes data collection, standardization procedure, and initial analysis. Chapter 4 is dedicated to examining publication bias. In Chapter 5, we explain the collected variables and examine the heterogeneity in our data set. In Chapter 6 we propose the best-practice estimate. Finally, Chapter 7 concludes the results.

Chapter 2

Unveiling the motivation

In this section, we highlight the importance of the topic from an economic perspective, discuss key concepts, and outline our contribution to the existing literature. Since one meta-analysis and multiple literature reviews have already been conducted, we only provide a brief overview of the relevant literature.

2.1 Relevance of the topic

Several studies have shown that changes in school organization can significantly impact academic performance (Betts 1998; Wößmann 2003). Consequently, identifying the positive peer effect of socioeconomic status (SES) may serve as a compelling rationale for implementing policy measures to address the resulting achievement gap (Paloyo 2020). Without such measures, students from high-SES backgrounds would benefit from attending school with their peers, while those from low-SES backgrounds may miss out on the beneficial influence of being in the company of high-SES peers. Furthermore, the educational system may fail to fully harness the potential of low-SES students. These inefficiencies could contribute to the exacerbation of inequality. A study of peer effects also yields insights for policy formulation aimed at promoting the success of disadvantaged minorities within society (Rangvid 2003; Berkowitz 2021).

Although the effect is subject to numerous academic disciplines, including Economics, Sociology, and Educational Psychology, we approach it primarily from an economic perspective. The topic is related to educational economics, specifically human capital production optimization. From this point of view, the overall economic objective is to maximize labor efficiency, subject to constraints. An example of a constraint is the school budget or teachers' limited capacity. The strategy of sorting students represents labor efficiency optimization. When assuming the presence of peer effects, ability grouping positively affects better-performing students and negatively affects low-performing students. On the other hand, establishing heterogeneous classes may benefit low-performing students and harm high-performing students.

How to optimally design an educational system depends on whether we aim to maximize the abilities of selected individuals or average ability. Put another way, whether we strive to build an outstanding elite or a more equitable society. If the social planner's objective was to maximize average academic achievement, the optimal policy is contingent upon the presence of either increasing or decreasing returns to peer groups (Rangvid 2003). Regardless of the goal or ideology, a thorough understanding of the effects is necessary to find the optimal solution. Therefore, we aim to elaborate on the effect and its drivers to enrich this debate. The significance of the peer effect has implications for the school choice debate and the ability grouping strategy. Both will be discussed in the following paragraphs.

An often debated question is whether students should be able to choose which school to attend instead of being required to attend the nearest school. The opponents of school choice fear it would make sorting by ability and social status more intensive. It could result in the exclusion of low-SES students from prestigious institutions, depriving them of beneficial peer effects. Conversely, school choice proponents question the existence and magnitude of peer effects or argue that the composition of students would remain unchanged from the status quo. Lastly, some note that sorting already happens when parents choose whether to live in a high-SES neighborhood or not.

A very similar discussion is being held at the group level. The fundamental question concerns whether to group students of similar abilities or establish heterogeneous classes. When assuming that the closest group impacts a student more than unknown students from the whole school, the debate about ability grouping in schools is even more pressing. The advocates of ability grouping in schools argue that clustering students of the same abilities enables lectors to adjust their teaching style to the needs of students, which would not be possible in more heterogeneous classes (Paloyo 2020). The opponents of ability grouping point out the positive effects that low-ability students gain when grouped with high-ability students, such as peer pressure and intellectual stimulation (Harel Ben Shahar 2022). Furthermore, it can be argued that average and above-average students do not benefit significantly from being grouped with peers of similar ability levels (Rangvid 2003).

Additionally, Parents and school principals are two key beneficiaries of information on true effect size. From the perspective of a school principal, the allure of being able to increase average school performance simply by reassigning students across different classrooms is quite tempting. On the other hand, as a parent, one can advocate for placing their child in an educational setting that enables them to reach their full potential or avoids exposing them to potentially detrimental environments (Paloyo 2020).

2.2 Fundamental concepts

Socioeconomic status

Understanding children's socioeconomic status (SES) became a primary concern for educational researchers after low academic performance was observed in children whose parents had low incomes, low levels of education, and low-status professions (Cowan *et al.* 2012). Since SES is one of the variables most commonly used in social science, it has been conceptualized in the literature in various ways (Rodriguez-Hernandez *et al.* 2020). A Frequently used definition by Mueller & Parcel (1981) describes socioeconomic status as *"the relative position of an individual or family within a hierarchical system, based on their level of access to, or control over, various highly valued commodities, including but not limited to wealth, power, and social status."* Recently, SES has been commonly defined as *"the amount of economic, social, and cultural resources available to a single student."* (De Clercq *et al.* 2017).

The method of measuring socioeconomic status is crucial, as different approaches yield different sizes of estimated effects (Sirin 2005). On the individual level, a consensus appears to exist regarding the three components proposed by Duncan *et al.* (1972), which comprise parental income, parental education, and parental occupation as the three primary indicators of SES. According to Erola *et al.* (2016), no matter when pupils are observed, education, occupation, and income can reliably reflect their socioeconomic status. On the family level, Sirin (2005) proposed household resources, such as books, computers, and designated study space, as a fourth indicator of SES.

Several studies, such as Harker & Tymms (2004) and Hutchison (2003), use the percentage of students eligible for free or reduced-price lunch (FLE) as a measure of peer socioeconomic status. The meta-analysis excludes studies that use free lunch eligibility due to the proposal of Hauser (1994) that researchers avoid using free lunch status as a variable when studying the impact of economic disadvantage. According to Hill & Jenkins (1999), a quarter of children aged 6-11 in the United Kingdom experienced poverty for 1-2 years between 1991 and 1996. Nonetheless, just 1.5% of the population suffered from poverty during the whole 6-year period. This instability contributes to the unreliability of FLE measures.

Academic achievement

Academic success holds significant value as a crucial personal and community asset and is linked to favorable outcomes. The lower academic achievement observed among vulnerable populations and students from disadvantaged backgrounds raises notable concerns among social work professionals and the general public (Berkowitz 2021). Poverty and privilege have a strong correlation with student's education, specifically in terms of literacy skills and reading abilities, leading to negative consequences for children and schools located in lower socioeconomic circumstances (Buckingham *et al.* 2013; Neuman & Celano 2015; Berkowitz 2022).

The notion of academic achievement (AA) is subject to various perspectives. This study included papers focusing on math, language, and science. As in the literature, we will use the terms "achievement," "success," and "performance" interchangeably. Even though solely focusing on academic performance may not adequately capture or reflect students' development of skills or resilience (York *et al.* 2015), we will adhere to academic achievement measures for the sake of simplicity.

Comparative testing appears to be one of the most reasonable ways to extract information on children's relative abilities. Since 2000, the OECD has conducted regular international comparative tests known as PISA (Programme for International Student Assessment). This meta-analysis encompasses seven OECD studies that have used data from these assessments. Additionally, many other included studies have used PISA data. PISA results are frequently used as they offer large sample sizes and various scarce variables such as parental involvement in education, parental academic interest, or social communication. Thus, when analyzing our data set, we ought to proceed with caution because many of the studies in this analysis used the same data and similar approaches.

Peer effects in education

The study of peer effects posits that children acquire knowledge and attitudes not only from teachers but also from their peers and the school environment. Schneeweis & Winter-Ebmer (2007) findings indicate that peer groups play a substantial role in providing inspiration and motivation. Part of this peer educational effect may be related to socioeconomic status, which we investigate in this thesis.

2.3 Literature overview

While some earlier studies have been published, the publication of the report by Coleman (1966) was one of the most pivotal moments in educational research during the twentieth century. This study, which included over half a million students, reported a strong correlation between socioeconomic status and academic achievement. Many other publications investigating this area have emerged in response to this paper. Studies from this period also frequently focused on race ratios in the classroom, which proved to be associated with differences in educational attainment. Indeed, they were partially intertwined with differences in economic status (Zimmer & Toma 2000).

In the following decade, several studies on peer effect and socioeconomic status emerged, including Henderson *et al.* (1978) and Summers & Wolfe (1977). However, studies provided inconclusive evidence. The methodology, which was the dominant contributor to variation, became the subject of discussion. For example, Evans *et al.* (1992) have argued that previous estimates of peer influence might suffer from bias as students choose peers. As White (1982) noted, using aggregated data on the school level for individual analysis can significantly influence the result. Those who do commit the so-called "ecological fallacy" (Borgatta & Jackson 1980; Robinson 2009). The term refers to a formal error in statistical data interpretation that arises when conclusions regarding individual characteristics are drawn from inferences made about the larger group or population to which these individuals belong.

There may be several problems when measuring peer effects. According to Rivkin (2001), the relationship between peer characteristics and student performance may not be causal because families with more resources or a greater commitment to education prefer to live in higher-income neighborhoods and attend schools with a higher-income student body. As a result, true peer effects can be confused with parental influences. This is in line with the findings of Jencks *et al.* (1990), who demonstrated that the size of estimated peer effects tends to decrease the more parental factors are controlled for. Moreover, several recent studies cast doubt on the existence of any link, direct or indirect, between socioeconomic status and academic achievement. For instance, Marks (2016) demonstrates that when early childhood cognitive ability and previous academic performance are taken into account, the effect diminishes. In the case of Chile, Gutiérrez (2023) discovered minimal effects associated with changes in the socioeconomic status of classmates.

Parallel to this, multiple scholars tried to investigate how the effect differs across the distribution of students (low to high performing). Most of them concluded that low-SES students benefit more from the improvement of their peer group, including Summers & Wolfe (1977), Zimmer & Toma (2000), and Rangvid (2003). According to Aram & Levin (2001), children at the lower end of the conditional achievement distribution are "dependent" learners compared to their more knowledgeable classmates, for whom the benefit of being around like-minded people is minimal. The amount of existing literature and the ambiguity of results provide good reason to perform a meta-analysis, as was carried out by Van Ewijk & Sleegers (2010).

2.4 Contribution

Our research builds upon the previous meta-analysis conducted by Van Ewijk & Sleegers (2010). Our contribution stems from the construction of a unique data set and employment of the most up-to-date statistical methods for meta-analysis in Economics, which are in line with guidelines provided by Havránek *et al.* (2020).

Previous meta-analysis (Van Ewijk & Sleegers 2010) addressed the issue of publication bias only by assessing the correlation between standard error, revealing little to no evidence of publication bias. Our research plan involves conducting an extensive investigation into publication bias by implementing a diverse range of statistical tests. Our objective is to clarify how this bias may potentially influence the behavior of the effect, a relationship that has not been adequately addressed in previous research.

Furthermore, a notable methodological contribution we aim to make is our rigorous exploration of heterogeneity within the existing literature. As we delve into a previously unexplored data set, we use model averaging techniques to quantify the impact of various factors on the underlying effect. By undertaking these measures, our thesis can serve as a robustness check for the existing findings applied within a distinct data environment. We compiled a unique data set comprising a large number of studies, notably the latest research from 2023. Additionally, we expanded the scope of variables collected compared to the previous study, resulting in a more comprehensive data set.

Chapter 3

Constructing the data set

3.1 Literature search

The initial step in constructing our data set involves thoroughly searching for scholarly studies that report on the studied effect. Google Scholar was used due to its robust full-text search functionality. The search query yielded over 16,000 results, of which the first 300 were examined. Figure A.1 in the Appendix outlines the specifics of the procedure, including the search query and Prisma diagram. Furthermore, we included studies from the previous meta-analysis (Van Ewijk & Sleegers 2010) that met the inclusion criteria. We also employed a technique called "snowballing". Snowballing is the practice of looking through study references to discover further estimates that are suitable. The search was completed on May 1, 2023.

A total number of 449 estimates from 40 studies that met all the specified inclusion criteria (described on the following page) were encoded and analyzed. Notably, our data set covers almost 40 years of research. An overview of the included studies can be found in Table A.1 in Appendix A. During the selection of studies and data collection, this study followed the guidelines for meta-analysis in economic research provided by Havránek *et al.* (2020).

To be included in the data set, the following requirements must be met:

1. The study must report standard errors or any other statistics that can be used to calculate standard errors. Many meta-analytic techniques require standard errors as weights, and we need them to determine the degree of publication bias.

2. The study must directly or indirectly estimate the impact of a one-standard-deviation increase in the average SES of the peer group, i.e. the children with whom a pupil attends school. Studies based solely on categorical variables (such as schools with a particular proportion of students from low-income families) are excluded because the effects of this type of variable cannot be reliably transformed into estimates of the selected type.

3. Individual student's educational achievement must be used as the dependent variable in the model, which is measured by scores on tests of mathematics, language, science, or a combination of these. Studies that only measure educational achievement in broad categories, such as passing or failing exams or dropping out of school, were excluded because they focus on a specific point in the distribution, namely the lower end. In contrast, our focus is on the overall distribution shift.

4. The individual-level SES variable must be included as a covariate in the estimation model. Failure to do so may result in the average SES variable being used as a proxy for individual student's SES, resulting in a significant overestimation of the peer effect due to the strong correlation between the two variables.

5. Students in the study sample must be within the primary and secondary (high) school age range.

We followed the reasoning outlined in the publications of Van Ewijk & Sleegers (2010) and Cala *et al.* (2022).

3.2 Variable selection

To determine which variables are most relevant to our research objectives, we read each of the selected studies in depth, as well as multiple other meta-analyses. A total of 48 aspects of data were collected for each study. An overview of all collected variables, including the rationale for the selection of variables, can be found in Chapter 5. Some variables were eliminated during the initial analysis phase because they were present in insufficient numbers of cases, while others were eliminated subsequently due to their high Variance inflation factor value. As a result, fewer variables are examined in model averaging.

In this thesis, we investigate the effect of increasing the average peer group's socioeconomic status by one individual-level standard deviation on achievement, measured in standard deviations. Thus, we collected the effect, its standard error, or t-statistics, as well as information on the methodology employed, such as the use of particular procedures or model specifications. We also encoded information on the characteristics of SES and Academic achievement variables. In addition, we encoded information about the sample, such as age, country, and level (class or school). Finally, we encoded variables regarding the relevance and impact of the study, including number of citations and publication status. Overall, more than 27,000 data points were obtained.

3.3 Comparability of effects

When needed, we linearly transformed the effects and standard errors to reflect the effect of a one-standard-deviation increase in SES on the AA variable's standard deviation. To do so, information on both the standard deviation of SES and the standard deviation of the academic achievement variable must be provided. Several papers have been eliminated because they failed to report these values.

It is necessary to explain why we employ an individual-level standard deviation rather than one standard deviation in the average SES distribution of the school (or class or cohort). A comparison of estimates from different populations (and consequently different levels of segregation) is problematic as the standard deviation of school average SES is dependent on the degree of segregation in a population. Using individual-level SES standard deviation results in more comparable estimates as it more accurately reflects population segregation differences. To further account for differences between nations, we encoded Gini coefficients based on the countries where the data was gathered.

3.4 Primary analysis

Some studies provide only one estimate, such as Young & Fraser (1993) with an estimate of 0.121. Other studies, including McEwan (2003) or OECD (2003) provide many estimates encompassing a broad range of values. As depicted in Figure 3.2, most studies report positive effects. However, the magnitude of the effect varies significantly both between and within studies, ranging from -0.278 (McEwan 2003) to 1.767 (Kartianom & Ndayizeye 2017). Neither the minimum nor the maximum estimate in our data set can be seen as reliable. It is essential to note that OECD-published studies provide, on average higher positive estimates and constitute a significant portion of our estimates. In some cases, weighting procedures are utilized to address this difficulty.

Table 3.1 provides the basic summary statistics of collected estimates and standard errors. The average estimate in our data set is 0.333. Nevertheless, this number is insufficient for drawing conclusions since a large proportion of estimates were derived from 7 OECD studies, which all use very similar methods. This notion is supported by the smaller value of 0.241 obtained when calculating the weighted average. The weights here correspond to the inverse of the number of estimates taken from each paper.

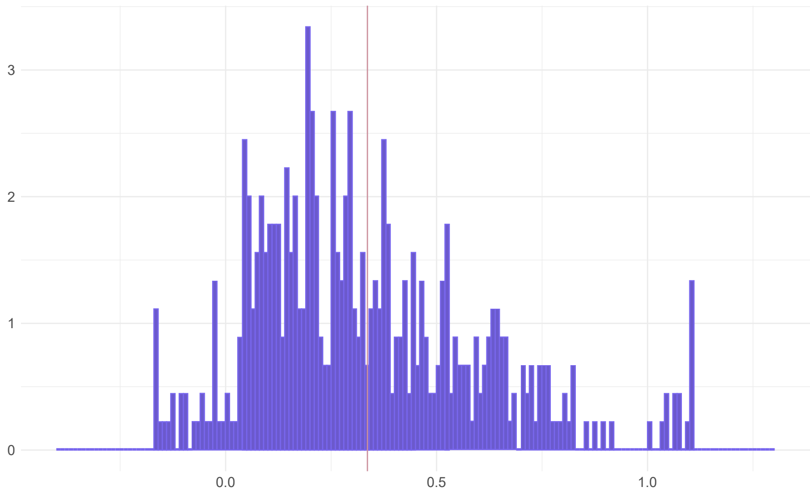
Table 3.1: Overview of Estimates and Standard Error

	Mean	Median	SE	Weighted mean
Estimate	0.333	0.286	0.276	0.241
Standard error	0.054	0.039	0.003	0.056

Notes: The results came from the entire sample. The weighted mean was calculated as the inverse of the number of estimates provided by that particular study. SE = standard error

Notably, implementing appropriate winsorization (1%) was deemed necessary due to the significant variability observed in our data set. The Winsorization technique is used to cope with outliers in data by reducing their degree of extremeness. Given the limited number of estimates available, this method is more appropriate than trimming the outliers. Nonetheless, leaving outliers untreated may distort the results of the analysis.

Figure 3.1: Distribution by effect magnitudute



Notes: The histogram of the estimates provides insight into the distribution of the effect sizes we collected. The sample mean is represented by the pink vertical line. Outliers are omitted from this histogram for conciseness but included in all statistical tests.

We proceeded to conduct a more in-depth analysis of the gathered data. The distribution of effect sizes is graphically represented in Figure 3.1. The effect sizes appear to be skewed to the right and most densely distributed between 0.1 and 0.4. It is worth noting that negative estimates are also evident in the distribution.

Furthermore, we present summary statistics for different subsamples in Table 3.2. This table allows us to examine how various characteristics of the studies influence the mean. We also provide the reader with a weighted mean to account for numerous OECD estimates, as explained earlier in this section. In order to observe differences between the lower and upper halves of continuous variables, the median value was used to divide them into two categories.

To comment on the obtained subsamples, we observe a higher average of 0.426 for studies conducted by the OECD. In addition, studies with smaller Study sizes (i.e., fewer estimates contributed) have a substantially lower sample mean. Again, this can be explained by OECD studies that provide numerous large estimates. Furthermore, published studies exhibit a sample mean of 0.246, whereas the overall mean across all studies is 0.333. Intuitively, we observe a significantly lower mean when we only look at studies that controlled for prior attainment in estimation regression or tried to overcome omitted variable bias. We also discovered an unexpectedly reduced average for the General academic achievement test category. Such a test type would presumably capture pupil differences more accurately, leading to higher estimates. The effect also appears to diminish with age. Notably, studies from the Social science field produced, on average, larger estimates than studies from Economics and Psychology.

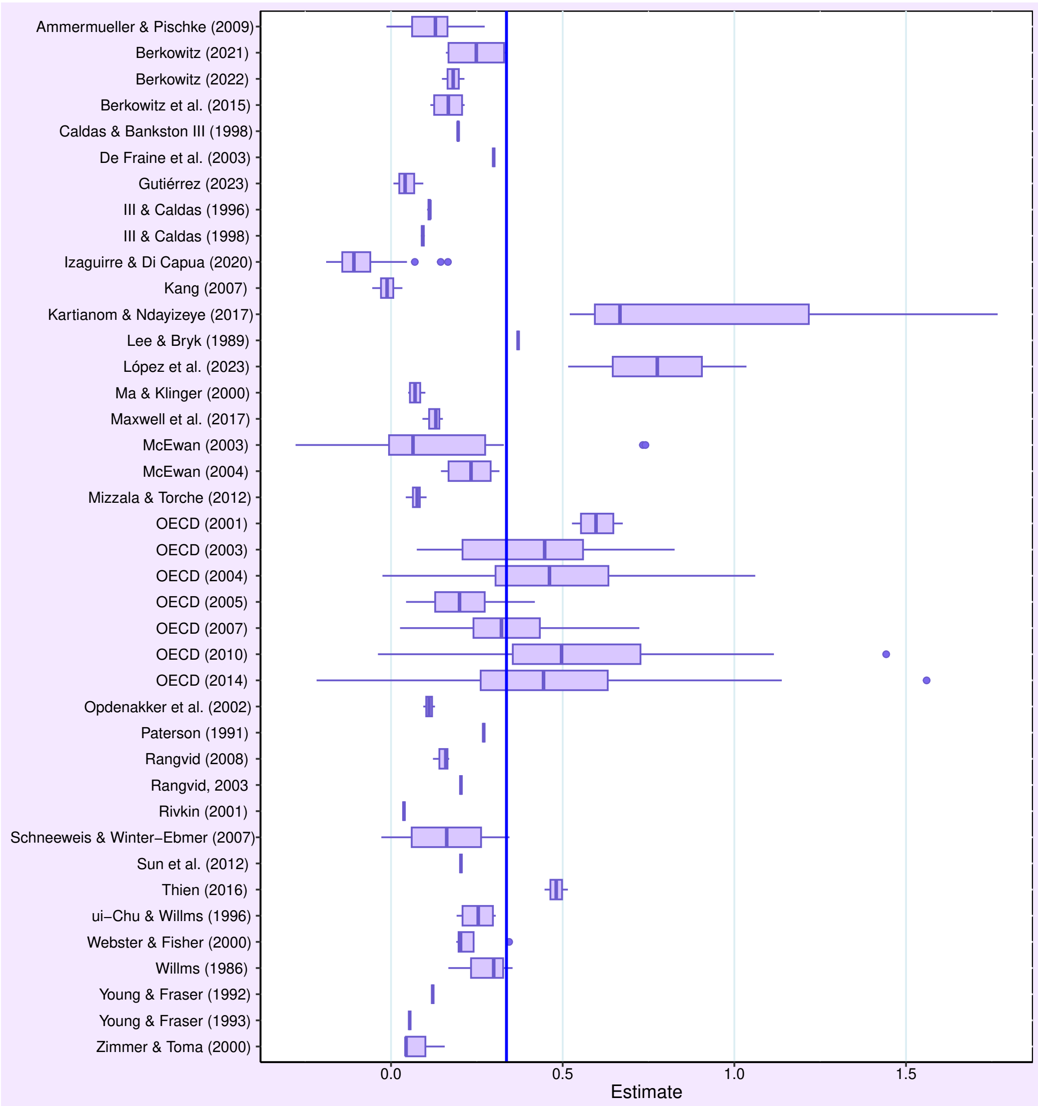
Table 3.2: Summary statistics of selected subsets

Variable	Sample Mean	CI	Weighted Mean	WM CI	Observations
All Data	0.333	$(-0.192; 0.858)$	0.241	$(-0.284; 0.766)$	449
Citations $\geq$ 317	0.334	$(-0.168; 0.836)$	0.224	$(-0.278; 0.726)$	229
Citations $<$ 317	0.333	$(-0.216; 0.882)$	0.249	$(-0.300; 0.798)$	220
Study size $\geq$ 39	0.454	$(-0.036; 0.944)$	0.449	$(-0.041; 0.939)$	258
Study size $<$ 39	0.171	$(-0.213; 0.555)$	0.211	$(-0.173; 0.595)$	191
If study was published	0.246	$(-0.293; 0.785)$	0.205	$(-0.334; 0.744)$	208
Publication Year $\geq$ 2007	0.359	$(-0.217; 0.935)$	0.270	$(-0.306; 0.846)$	257
Publication Year $<$ 2007	0.299	$(-0.138; 0.736)$	0.217	$(-0.220; 0.654)$	192
T-statistic $\geq$ 6.847	0.488	$(0.018; 0.958)$	0.422	$(-0.048; 0.892)$	225
T-statistic $<$ 6.847	0.178	$(-0.202; 0.558)$	0.163	$(-0.217; 0.543)$	224
AA Language	0.332	$(-0.193; 0.857)$	0.239	$(-0.286; 0.764)$	201
AA Math	0.370	$(-0.194; 0.934)$	0.301	$(-0.263; 0.865)$	166
AA Sciencee	0.288	$(-0.128; 0.704)$	0.190	$(-0.226; 0.606)$	69
AA GAA	0.113	$(-0.034; 0.260)$	0.126	$(-0.021; 0.273)$	12
SES Class Level	0.145	$(-0.249; 0.539)$	0.182	$(-0.212; 0.576)$	57
SES Home Resources	0.422	$(-0.103; 0.947)$	0.290	$(-0.235; 0.815)$	244
SES Parental Income	0.370	$(-0.181; 0.921)$	0.260	$(-0.291; 0.811)$	319
SES Parental Occupation	0.393	$(-0.107; 0.893)$	0.305	$(-0.195; 0.805)$	359
SES Dichotomous	-0.015	$(-0.268; 0.238)$	0.030	$(-0.223; 0.283)$	27
SES Composite	0.381	$(-0.172; 0.934)$	0.256	$(-0.297; 0.809)$	297
SES Combined	0.377	$(-0.162; 0.916)$	0.269	$(-0.270; 0.808)$	357
Average SES $>$ 1	0.058	$(-0.312; 0.428)$	0.067	$(-0.303; 0.437)$	55
Number of obs. $\geq$ 5796	0.285	$(-0.274; 0.844)$	0.231	$(-0.328; 0.790)$	225
Number of obs. $<$ 5796	0.382	$(-0.086; 0.850)$	0.256	$(-0.212; 0.724)$	224
Average Student Age $\geq$ 15	0.413	$(-0.079; 0.905)$	0.318	$(-0.174; 0.810)$	328
Average Student Age $<$ 15	0.118	$(-0.229; 0.465)$	0.166	$(-0.181; 0.513)$	121
Discipline: Economics	0.091	$(-0.250; 0.432)$	0.105	$(-0.236; 0.446)$	79
Discipline: Social Science	0.388	$(-0.114; 0.890)$	0.291	$(-0.211; 0.793)$	366
Discipline: Psychology	0.124	$(0.065; 0.183)$	0.124	$(0.065; 0.183)$	3
Country: Asia	0.373	$(-0.193; 0.939)$	0.215	$(-0.351; 0.781)$	57
Country: South AM	0.176	$(-0.304; 0.656)$	0.209	$(-0.271; 0.689)$	93
Coutry: North AM	0.314	$(-0.198; 0.826)$	0.214	$(-0.298; 0.726)$	35
Country: Europe	0.402	$(-0.113; 0.917)$	0.253	$(-0.262; 0.768)$	197
GINI coefficient $\geq$ 34.7	0.290	$(-0.212; 0.792)$	0.262	$(-0.240; 0.764)$	226
GINI coefficient $<$ 34.7	0.377	$(-0.158; 0.912)$	0.217	$(-0.318; 0.752)$	223
Attempt to Overcome OVB	0.154	$(-0.199; 0.507)$	0.152	$(-0.201; 0.505)$	69
Prior attainment control	0.157	$(-0.245; 0.559)$	0.198	$(-0.204; 0.600)$	29
Primary concern	0.161	$(-0.196; 0.518)$	0.190	$(-0.167; 0.547)$	79
PISA data	0.422	$(-0.070; 0.914)$	0.386	$(-0.106; 0.878)$	316
OECD	0.426	$(-0.062; 0.914)$	0.436	$(-0.052; 0.924)$	302

Notes: The table provides summary statistics for selected subsets. Weighted refers to weighting the estimates based on the inverse number of estimates provided by each study. CI = confidence interval, WM CI indicates confidence intervals for weighted means, and Observations indicate the number of observations. See table 5.1 for a detailed explanation of the variables. The subsamples were divided by the median value.

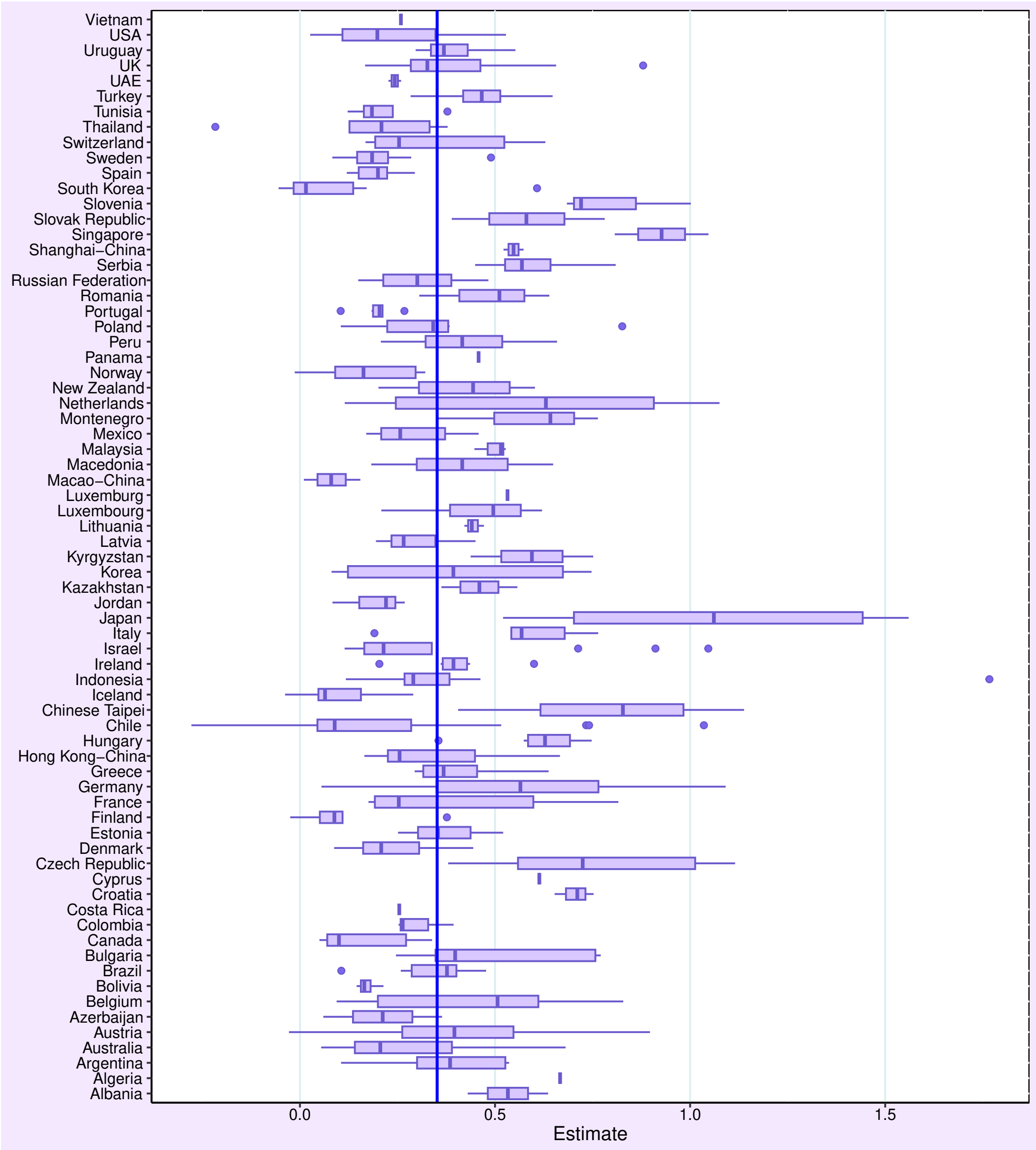
To visually illustrate the variation in estimates among specific studies and their reported effects, we present Figure 3.2. This visualization highlights the significant discrepancies in estimates depending on the study author. Finally, Figure 3.3 displays the substantial divergence in estimates based on the country where the data was collected. We omitted estimates that were derived from multiple countries simultaneously from the graph.

Figure 3.2: Effect size across individual studies



Notes: The figure shows a box plot of the recalculated estimates from individual studies. The dividing line inside each box represents the median value, and the length of each box represents the interquartile range. Whiskers cover (P25 - 1.5* interquartile range) to (P75 + 1.5* interquartile range). The blue vertical line represents the mean.

Figure 3.3: Efect size across countries



Notes: The figure shows a box plot of the recalculated estimates grouped by countries. The dividing line inside each box represents the median value, and the length of each box represents the interquartile range. Whiskers cover (P25 - 1.5* interquartile range) to (P75 + 1.5* interquartile range). The blue vertical line represents the sample mean.

Chapter 4

Publication bias

Drawing from intuition and existing literature, one would anticipate positive estimates of the effect of peer socioeconomic status (SES) on academic achievement. Consequently, disregarding zero or negative estimates may be perceived as logical. In light of this logic, a novel field of investigation has emerged, addressing the phenomenon of publication bias. Previous studies, notably the work of Ioannidis *et al.* (2017), revealed that this practice of discarding unexpected estimates distorts the conclusions drawn from the existing body of research. Nevertheless, publication bias in the economic literature is an inherent and unavoidable phenomenon that does not imply any intentional actions by authors, editors, or reviewers. It is the responsibility of those who evaluate and analyze the existing body of literature to address this bias (Cala *et al.* 2022).

Publication bias arises from two main sources. The first issue, commonly known as the "file-drawer problem" (Stanley 2005), refers to the idea that less significant findings tend to be "left in the drawer" and remain unpublished. Simultaneously, significant findings are more likely to be published. The second source of publication bias is attributed to individual scientists. Even when the true underlying effect is consistently positive, the presence of noise in the data and methodologies can lead to both negative and statistically insignificant (zero) estimates. However, researchers are more inclined to focus on specifications that yield apparent positive effects, as they are believed to be closer to the truth. Another problem arises when the noise in the data and methodologies generates estimates significantly larger than the true effect. Identifying such implausible estimates is challenging because no upper threshold symmetrical to zero would alert the researcher to their implausibility. Consequently, an upward bias is introduced if numerous small and imprecise estimates are disregarded while large and imprecise estimates are reported. Hence, a paradox emerges: while publication bias may offer benefits at the individual study level, it proves detrimental when considering the collective body of literature as a whole (Gechert *et al.* 2022).

This study closely followed methodology of Havranek *et al.* (2021) and Gechert *et al.* (2022), when searching for publication bias[1]. It is also important to draw attention to Aguinis *et al.* (2011), who highlighted the potential limitations of certain methods used to test for publication bias, cautioning that these approaches may yield misleading results due to their reliance on a limited amount of information.

Regarding the peer effect of SES, we anticipate that the publication bias will introduce a positive upward bias. The prior meta-analysis by Van Ewijk & Sleegers (2010) addressed publication bias solely by calculating the correlation between standard error and effects. The authors argued that the observed correlation could be attributed to one misleading data point. They contend that by excluding this particular data point from the analysis, the presence of publication bias is no longer evident. Thus, we find it crucial to examine the previously unexplored topic of potential publication bias in this field.

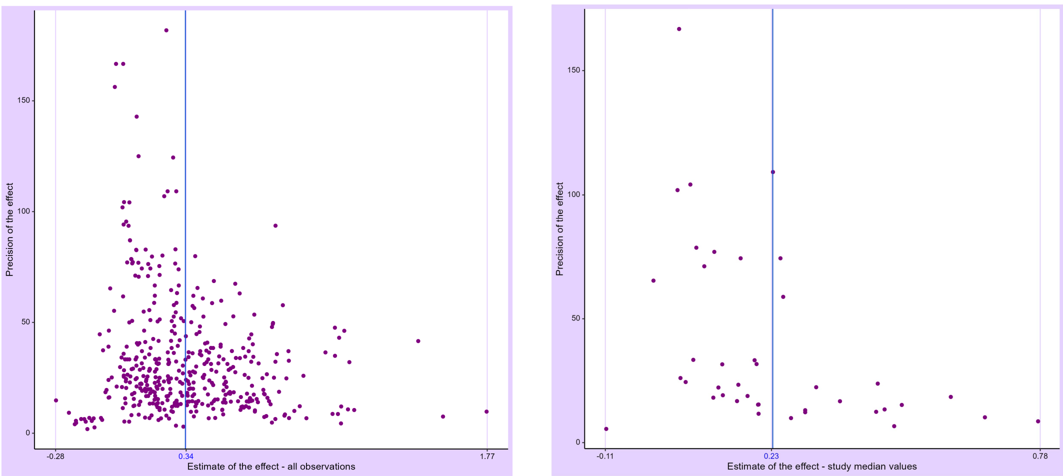
[1]Among the other studies on publication bias in economics are Stanley (2001), Iršová *et al.* (2010), Babecky & Havranek (2014), Havranek *et al.* (2017), Havranek *et al.* (2018), Havranek & Sokolova (2020), Havránek *et al.* (2020), Bajzik *et al.* (2020) Havranek *et al.* (2022), Elminejad *et al.* (2022).

4.1 Funnel plot

We start with constructing The Funnel plot proposed by Egger *et al.* (1997) as it is one of the most prominent visual methods for detecting publication bias. In Figure 4.1, the estimates are plotted against their corresponding precision, represented by the inverse of the standard error. In the absence of publication bias, the scatter plot will form an inverted funnel: the most precise estimates will lay close to the true underlying effect, while imprecise estimates will be more dispersed and symmetrically distributed around the true effect (Gechert *et al.* 2022). An asymmetry of the funnel plot indicates publication bias, and the top of the funnel indicates the underlying mean effect adjusted for the bias. The evidence in Figure 4.1 is consistent with possible discrimination against estimates with a counterintuitive negative sign, as we observe asymmetry. The most precise estimates are centered in the area between 0 and 0.25. The asymmetry becomes even more evident when looking solely at the study median values. To conduct a more thorough analysis of selection bias, we implemented additional statistical tests, which will be discussed in the following sections.

Figure 4.1: Funnel plot

(a) all observations (b) study median values



Notes: Funnel plot proposed by Egger *et al.* (1997). In the absence of publication bias, the plot should be symmetrical. Plot (a) depicts all observations, while plot (b) displays only the study median values. The blue vertical line represents the sample mean, and the mean of study medians, respectively. SE = standard error

4.2 Linear tests for selective reporting

In order to delve deeper into the examination of potential selective reporting, we conduct the Funnel Asymmetry Test (FAT) and the Precision Effect Test (PET). These tests investigate potential associations between the estimates and their corresponding standard errors through simple regression analysis. According to theory, the estimates of the effect should be randomly distributed around the mean estimate of the peer effect (Havranek *et al.* 2015). Nevertheless, if some results are preferred over others, the reported estimates of the effect will be correlated with their standard errors (Card & Krueger 1995).

Assuming that publication bias can be modeled as a linear function of the standard error and absence of heterogeneity across the studies, the intercept represents the mean estimate adjusted to account for the impact of standard errors. Hence, in this context, the intercept can be interpreted as the "true effect" (Stanley 2005). Nevertheless, it is essential to note that the linearity assumption may not hold universally, as discussed by Andrews & Kasy (2019). This matter will be addressed later in the thesis.

To perform the regression analysis, the following equation is estimated:

$$\text{effect}_{ij} = \text{effect}_0 + \beta_1 * (SE_{\text{effect}})_{ij} + \text{u}_{ij} \tag{4.1}$$

The effect_{ij} represents the i-th estimate along with its corresponding standard error $(SE_{\text{effect}})_{ij}$, obtained from the j-th study. The effect_0 denotes the "true effect", and u_{ij} denotes the error term. In the following tables, the term "effect beyond bias" is used to denote effect_0, while "publication bias" refers to β_1.

Table 4.1 presents the outcomes of various specifications derived from Equation (4.1), following the methodological framework inspired by prominent scholars such as Stanley (2008), Stanley *et al.* (2013), Stanley & Doucouliagos (2015). Unless explicitly mentioned otherwise, we employ clustered standard errors at the study level and assume exogeneity within the model.

Table 4.1: Linear tests for selective reporting

	OLS	**FE**	**BE**	**RE**	**Precision**	**Study**
Publication bias	0.063	4.312***	1.709**	1.314***	4.319***	0.882*
Standard error	(0.409)	(0.077)	(0.878)	(0.35)	(0.669)	(0.494)
Effect beyond bias	0.33***	0.166***	0.148***	0.176***	0.165***	0.391***
Constant	(0.021)	(0.002)	(0.056)	(0.039)	(0.026)	(0.028)

Notes: The table shows the results of estimation equation 4.1. OLS = ordinary least squares, FE = study-level fixed effects, BE = study-level between effects, RE = study-level random effects, Precision = estimates are weighted by the inverse of their standard error, Study = estimates are weighted by the inverse of the number of observations reported per study. *, **, and *** indicate statistical significance at 10%, 5% and 1% level. Standard errors are in parentheses.

The first column of Table 4.1 presents the findings of a simple OLS regression. The second column incorporates study-level fixed effects (FE) to control for unobserved study-specific characteristics. The third specification employs between-study (BE) variance instead of within-study variance. In the fourth specification, random effects (RE) are used, accounting for heterogeneity between studies by weighing both within-study and between-study variance (Bom & Rachinger 2019).

Subsequently, two weighting schemes are applied. First, precision (the inverse of standard error) is used as a weight, as suggested by Stanley & Doucouliagos (2017), addressing heteroscedasticity. Secondly, in order to ensure equal influence from each study on the final outcome, the data is weighted by the inverse of the number of estimates collected by each study. The corresponding estimate of the underlying effect raises concerns due to the fact that all seven OECD (2001; 2003; 2004; 2005; 2007; 2010; 2014) studies use a very similar approach, which tends to produce on average substantially higher estimates than other studies. Consequently, the average effect derived from this method reflects the OECD approach much more than other approaches. It is wise to interpret this estimate with caution and refrain from placing excessive weight on its implications. Notably, such a high constant may indicate that only a few studies drive the publication bias.

Overall, the findings of the FAT-PET tests suggest that the true effect size might be smaller than what is typically reported. In the majority of the specifications, we observe positive and statistically significant results, indicating the presence of publication bias. Additionally, we find a statistically significant and positive intercept, representing the bias-corrected mean effect. Following the correction for publication selection, the estimated mean effect appears to be, on average, around 0.2. This is equivalent to a one standard deviation increase in peer socioeconomic resulting in a 0.2 standard deviation increase in academic achievement outcomes for the individual.

4.3 Non-linear tests for selective reporting

Although the tests from the previous chapter serve as a solid foundation for detecting publication bias, they rely on the assumption of a linear relationship between the effect size and corresponding standard error. This assumption results in an imprecise estimation of publication bias if the relationship is non-linear or exhibits "jumps" around crucial values. Notably, the FAT-PET method tends to underestimate the "true effect" when being other than zero (Stanley *et al.* 2013; Bom & Rachinger 2019).

In reality, publication bias is unlikely to affect estimates with sufficient precision to achieve statistical significance at or below the 5% level. In such situations, a linear approximation would overcorrect for publication selection and introduce a downward bias, distorting the results in the opposite direction. We address this issue by using alternative approaches that allow for a non-linear relationship between publication bias and standard errors. The outcomes of applying six distinct methods are shown in Table 4.2.

Table 4.2: Non-linear tests for publication bias

	Top10	**WAAP**	**Stem**	**HBM**	**SM**	**EK**
Publication bias	-	-	-	0.704*	0.275***	4.312**
Standard error	-	-	-	(0.434)	(0.014)	(0.673)
Effect beyond bias	0.171***	0.338***	0.117	0.21***	0.279***	0.166***
Standard error	(0.022)	(0.012)	(0.149)	(0.074)	(0.021)	(0.017)

Notes: Top10 = method using only the top 10% most precise estimates (Stanley *et al.* 2010), WAAP = Weighted average of adequately powered estimates (Ioannidis *et al.* 2017), Stem = the stem-based method (Furukawa 2019), HBM = Hierarchical Bayes model (Allenby & Rossi 2006), SM = the Selection model (Andrews & Kasy 2019) where P represents the probability that estimates lacking statistical significance at the 5% level are published relative to the probability that statistically significant estimates are published (with the latter being normalized to a value of 1), EK = the Endogenous kink method (Bom & Rachinger 2019). *, **, *** indicate statistical significance at 10%, 5% and 1% level. Standard errors are in parentheses.

We commence with the "Top 10" approach introduced by Stanley *et al.* (2010). This method involves calculating the simple average of the 10% most precise estimates, which has been found to significantly mitigate publication bias and often yield more efficient estimates of the underlying effect. It is worth noting, however, that the "Top 10" approach conflicts with the Central Limit Theorem. The method produced an average effect estimate of 0.171. This estimate, compared to the overall data set's average of 0.333, indicates the presence of publication bias in the analyzed data.

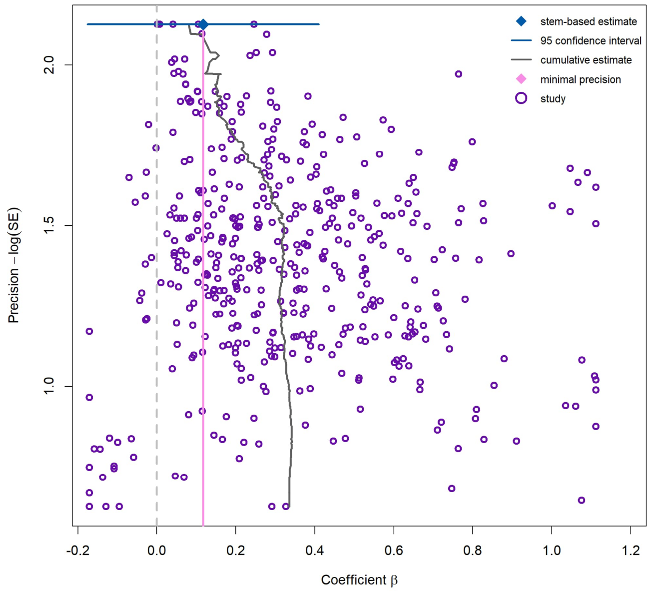
The Weighted Average of Adequately Powered (WAAP) technique proposed by Ioannidis *et al.* (2017) acknowledges the prevalent tendency to publish estimates based on their attainment of statistical significance, i.e., by exceeding the conventional threshold of 1.96 t-statistic. The authors propose employing Unrestricted Weighted Least Squares (WLS) exclusively for estimates derived from studies with adequate statistical power. This condition is evaluated by comparing the calculated standard errors to a power threshold determined by both statistical significance and adequate power (1.96 + 0.84, where the former is derived from statistical significance and the latter is derived from the definition of adequate power). Notably, the criterion is satisfied for over 130 estimates within our data set. This method produces a corrected effect estimate of 0.338, which is greater than the average value of the data set. Thus, the results do not indicate the existence of publication bias.

Furthermore, we employ the stem-based bias correction method introduced by Furukawa (2019). This method recommends using a subset of the most accurate estimates, referred to as the stem of the funnel plot. This subset is selected based on the optimal trade-off between reducing variance (achieved by including fewer estimates) and increasing bias (resulting from including more imprecise estimates). The stem-based approach tends to be more cautious than commonly employed methods, leading to wider confidence intervals. The stem-based method yields an estimated coefficient of 0.117, suggesting a strong selection bias.

Next, we apply the Hierarchical Bayes model by Allenby & Rossi (2006). This method employs Bayesian statistics and leverages within-study variation to determine the weights assigned to individual estimates, which are then pooled at the study level. For additional information, please see the original paper. This method returned an estimate of effect beyond bias of 0.21.

We extend our investigation of publication bias by employing the Selection model introduced by Andrews & Kasy (2019). This model proposes a correc-

Figure 4.2: Stem based method



Notes: The figure depicts an estimate of the "true effect" from Stem-based method by Furukawa (2019). The blue diamond represents the estimate of the true effect, the blue line represents the 95% confidence interval. The dark gray line represents estimates at various levels of precision. The violet circles represent individual estimates of the effect. The logarithm was used to reduce the disparity between standard error values.

tion for publication bias by utilizing the "conditional publication probability." This probability reflects the likelihood of a study being published based on its obtained results.

Finally, we utilize the Endogenous Kink (EK) meta-regression model proposed by Bom & Rachinger (2019). This model incorporates the identification of a kink at a specific cutoff value of the standard error. Below this threshold, the occurrence of publication bias becomes highly improbable. Once this kink is determined, Bom & Rachinger (2019) suggests fitting a piecewise linear regression of the collected estimates on their corresponding standard errors to uncover the true effect.

Considering the combined results of the nonlinear models, there is compelling evidence that the corrected peer effect on academic achievement associated with socioeconomic status is approximately 0.2. It is worth noting that the uncorrected mean estimate of 0.333 indicates a significant exaggeration due to publication bias. However, the magnitude of publication bias does not appear to be twofold or more, thus our findings are not fully consistent with the rule of thumb put forth by Ioannidis *et al.* (2017).

4.4 Extensions

To enhance the robustness of our findings, we decide to relax the previously held exogeneity assumption. Thus, even when publication bias is not present, we allow for a correlation between standard errors and the studied effect. Endogeneity in the standard error can arise from three potential sources. Firstly, a measurement error occurs as the standard error itself is an estimate. Secondly, reverse causality can occur when researchers manipulate the standard error, either intentionally or unintentionally, to obtain statistically significant estimates. Lastly, unobserved heterogeneity may introduce a systematic influence on both estimates and standard errors through methodological choices.

We commence with an examination of the potential endogeneity of standard error using an Instrumental Variable (IV) regression. Using the inverse of the square root of the degrees of freedom as the instrument for the standard error turned out to be the most appropriate approach. This instrument maintains an inherent correlation with the standard error but mitigates the aforementioned sources of endogeneity (Havranek *et al.* 2022). However, even the most suitable instrument demonstrates weakness, as the first stage F-statistics is well below any significance threshold (0.272). The results can be seen in Table 4.3.

Table 4.3: Adressing potential endogeneity

	IV	p-uniform*
Publication bias	-61.38	L=0.536
Standard error	(213.27)	(p=0.464)
Effect beyond bias	3.59	0.333
Standard error	(11.184)	(0.475)

Notes: IV = Instrumental Variable Regression, where the instrument is the inverse of the square root of the number of observations. Standard errors, clustered by study, can be found in parentheses. None of the values are statistically significant. P-uniform* introduced by van Aert & Van Assen (2021), where L is the test statistic of p-uniform*'s publication bias test and the associated p-value is presented in brackets.

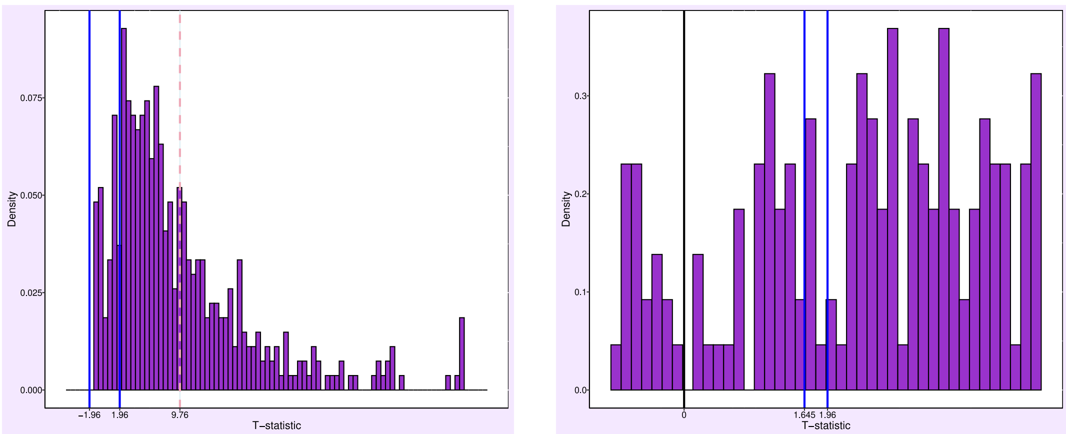
We proceed to the p-uniform* technique by van Aert & Van Assen (2021). The idea behind p-uniform* is that p-values should exhibit a uniform distribution around the true effect size when testing the hypothesis that the estimated coefficient equals the underlying value of the effect. Publication bias can influence specific segments of the p-value distribution, leading to an underrepresentation of large p-values and an overrepresentation of p-values slightly below the conventional significance level of 0.05. Applying p-uniform* using the maximum likelihood function results in insignificant results (see Table 4.3).

We continue our investigation of publication bias by performing the Caliper test proposed by Gerber *et al.* (2008). Unlike prior methods, this test makes no assumptions about the relationship between the effect and corresponding standard error. It instead focuses on the distribution of t-statistics. Specifically, Gerber *et al.* (2008) propose examining the distribution around specific statistically significant values, considering small intervals. The objective is to identify potential jumps in the distribution that may indicate over-reporting of a particular statistical value in the sample. If there is a tendency for any specific statistical value to be over-reported, it will be evident through a noticeable jump in the distribution. The density of t-values, along with relevant thresholds, is depicted in Figure 4.3. The data appears to be skewed to the right. Therefore, we further examine only positive significance thresholds. The Figure provides motivation to examine potential spikes rigorously.

Figure 4.3: Density of t-values

(a) overall density

(b) density around relevant thresholds



Notes: Figure (a) depicts the density of t-statistics. A visual inspection of the figure provides motivation to examine potential spikes further. The blue vertical lines correspond to the thresholds of -1.96 and 1.96 (indicating significance at the 5% level). The pink dashed line represents the mean value in the data set. The data is skewed to the right, so we continue to inspect only positive values. Figure (b) shows the density of t-statistics around relevant significance thresholds. The black vertical line represents 0, and the blue vertical lines correspond to thresholds of 1.645 and 1.96 (significance levels of 10% and 5%).

The results of the Caliper test can be found in Table 4.4. The obtained values can be interpreted as the disparity between the number of observations above and below a specific threshold. For instance, a coefficient of 0.31 would imply that 81% of the estimates surpass the threshold, while the remaining 19% fall below it. In our analysis, we investigate thresholds of 1.645 and 1.96, corresponding to the 10% and 5% significance levels, respectively. However, due to the limited number of data points around 0 and -1.96 thresholds, conducting similar tests is not feasible.

Due to the small number of observations, the calipers must be set sufficiently large to ensure statistical significance. Conversely, if the caliper is excessively wide, it may fail to effectively capture potential jumps around the threshold since values further away from the thresholds are less prone to bias. In our study, the narrowest caliper we employed consisted of only 16 observations. All tests performed suggest the presence of publication bias.

Table 4.4: Caliper tests for publiaction bias

Threshold	**1.645**	**1.96**
Caliper width 0.4		
Estimate	0.316***	0.250***
Standard error	(0.064)	(0.058)
Observations	20	16
Caliper width 0.6		
Estimate	0.300***	0.296***
Standard error	(0.048)	(0.032)
Observations	31	33
Caliper width 0.8		
Estimate	0.325***	0.302***
Standard error	(0.0363)	(0.027)
Observations	41	45

Notes: The table presents the findings of the Caliper test (Gerber *et al.* 2008). This test examines the proportion of estimates above and below a significant threshold for the t-statistic, in this case for 10% (1.645) and 5% (1.96) significance levels. Different caliper sizes are employed based on the number of available observations. The standard errors provided in parentheses are clustered at the study level. *, **, and *** indicate statistical significance at 10%, 5% and 1% level.

Table 4.5: P-hacking tests

A. P-hacking tests by Elliott et al. *(2022)*		
	Test for non-increasingness	**Test for monotonicity and bounds**
P-value	0.248	0.319
Observations ($p \leq 0.1$)	392	392
Total observations	459	459
B. MAIVE estimator by Irsova et al. *(2023)*		
	MAIVE coefficient	**F-test**
Coefficient	0.548***	5.007
Standard Error	(0.077)	

Notes: Panel A presents the results of p-hacking tests conducted by Elliott *et al.* (2022), including the histogram-based test for non-increasingness and the histogram-based test for monotonicity and bounds. Panel B reports the outcomes of the spurious precision robust approach using the MAIVE estimator proposed by Irsova *et al.* (2023). The F-test represents the test statistic for the instrumental variable (IV) first-step F-test. Cluster-robust standard errors are employed in the MAIVE estimation, as indicated within parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

We now present the outcomes of tests proposed by Elliott *et al.* (2022). These rigorous tests offer a significant advantage as they eliminate the need for a predetermined threshold of the t-statistic. Instead, they examine the entire distribution of p-values to assess publication bias. Notably, Havranek *et al.* (2021) have emphasized that these tests require a considerable amount of data to ensure robustness and accuracy. Regrettably, our study lacks a significant number of observations, leading to a reduced level of reliability in our findings. The results can be seen in Panel A of Table 4.5. We present the outcomes of the results of the histogram-based test for non-increasingness and the histogram-based test for monotonicity and bounds. The method did not detect significant publication bias.

Finally, in panel B of Table 4.5 we present the findings derived from implementing the spurious precision robust approach, utilizing the Meta-Analysis Instrumental Variable Estimator (MAIVE) developed by Irsova *et al.* (2023). It is an extension of the funnel plot models pioneered by Egger *et al.* (1997), Stanley (2005), and Stanley (2008). This test aims to address the issue that arises when meta-analysis gives greater weight to studies with lower standard errors, resulting in estimates that may already be biased. The researcher does not have access to the true precision and must estimate it, which opens the possibility of p-hacking to achieve statistical significance. P-hacking standard errors, such as through inappropriate clustering, introduces a potential bias in estimating the overall mean effect. Furthermore, specific methodological decisions can collectively affect both the estimates and their corresponding standard errors, thereby compromising the reliability of the conventional publication bias test.

To address these issues, one simple solution is to use the inverse of the square root of the sample size as an instrumental variable for the reported standard error. The inherent relationship between sample size and standard error justifies this approach, and it is difficult to artificially inflate the sample size through p-hacking (Opatrny *et al.* 2023). The first-step F-test is a test statistic for the instrumental variable (IV), which allows us to evaluate the instrument's validity. Regrettably, the instrument employed in our case exhibits weakness, thereby offering only limited information. The obtained MAIVE coefficient indicates a substantially high bias-corrected estimate of 0.548.

Overall, the results have indicated the presence of publication bias across the majority of employed methods. However, the outcomes concerning the mean estimate adjusted for publication bias exhibit a notable lack of consistency. Even though the findings provide evidence of a positive peer effect of socioeconomic status on student achievement, the estimated effect sizes range from 0.148 to 0.391 (considering statistically significant results). On average, the tests suggest the true effect to be around 0.2. Our findings are in contrast with the previous meta-analysis by Van Ewijk & Sleegers (2010), which found little to no publication bias.

Chapter 5

Heterogeneity

A study by Paloyo (2020) concluded that peer effects depend largely on the context. However, we have not accounted for the various study contexts thus far. In order to enhance our understanding of the specific determinants driving the effect of peer socioeconomic status, we proceed to examine heterogeneity within and among the primary studies. Specifically, we delve into the collected variables, considering the potential influence of their selection on the resulting effect and investigating their behavioral patterns as suggested by previous research. Then, we employ model averaging methods using these variables to address model uncertainty and establish a robustness check. The results will be further used to derive a best-practice estimate.

5.1 Variable overview

We commence with an overview of encoded variables and the rationale behind their choice. Even though we collected more aspects of data, we focus on 41 variables that may be related to the effect size and sign. In some cases, we do not employ all of the variables in the model to avoid a dummy variable trap. Notably, variables with a VIF above 10 were omitted from the analysis, following the conventional approach in meta-analysis. The overview of the variables, along with their description, mean, and standard deviation can be found in Table 5.1. Our selection of variables was mainly based on the differences observed among primary studies, whereas for more technical variables, we reviewed prior meta-analyses (for example, Havranek *et al.* (2021; 2022); Iršová *et al.* (2010)). For clarity, we categorize the variables as follows: SES measurement, AA test type, Sample characteristics, Estimation approach, and Publication characteristics.

Socioeconomic status measurement

As discussed in Section 2.2.1, there is no consensus regarding the optimal selection of variables that effectively capture socioeconomic status (SES). Consequently, SES is measured in various ways, and as highlighted by Sirin (2005), different approaches to measuring SES yield different effect sizes. This finding comes from the meta-analysis conducted on the direct effect of socioeconomic status on academic achievement. However, we anticipate encountering a similar situation when examining the peer effect of SES. Therefore, our objective was to comprehensively capture the diverse methodologies employed for measuring socioeconomic status.

We began by encoding the aspect that the primary study used to indicate the socioeconomic position of an individual. We focused on the four most prevalent indicators, proposed by Duncan *et al.* (1972) and Sirin (2005). The most frequently used single metric in assessing SES is Parental Occupation, followed by Parental Education, Home Resources, and Parental Income. Researchers often combine multiple indicators to address the limitations of employing a single approach. We encoded two distinct variables: Combined and Composite. The Combined variable denotes the combination of multiple indicators, even if they are from the same category (for example, the education of the mother and father). The latter variable is a subset of the former and represents a combination of two or more previously mentioned variables: Home Resources, Parental Education, Income, and Parental Occupation.

Additionally, we collected a variable indicating whether the SES measurement was dichotomous. It is assumed that dichotomously measured SES is less reliable. Notably, studies utilizing Free Lunch Eligibility were completely excluded, as explained in Section 2.2. Furthermore, we encoded information regarding the level at which average SES was calculated, whether at the class, school, or cohort level. It is hypothesized that measuring SES at the class level results in more pronounced effects, given that the influence of one's immediate peers is deemed particularly influential. Finally, we collected individual-level SES standard deviations to perform effect recalculations and obtain standardized values.

Academic achievement test type

Another element where inputs differ is the measurement of academic achievement. Thus, we gathered information on the type of tests administered, encompassing Language, Math, Science, and General academic achievement (GAA) tests. Notably, previous meta-analysis (Van Ewijk & Sleegers 2010) has shown no significant variations in outcomes based on the selection of test types. Additionally, we collected information on the standard deviation of each academic achievement test to perform recalculations of the effect.

Sample characteristics

This category includes information on the characteristics of each sample. We collected information on sample size, which might potentially be used as a proxy for precision. In addition, the average age of the students in the sample was encoded because the degree of the effect may change with age. It is hypothesized that as children get older, the impact of adults like parents and teachers on their behavior will lessen, while the influence of their peers their age will increase (Van Ewijk & Sleegers 2010).

Figure 3.4 demonstrates that another substantial source of heterogeneity in the obtained results is the geographical region where the sample was taken. As a result, we gathered information on the origin of each estimate. To determine whether an estimate came from one country, we used a dummy variable called Country-level Data. We added the corresponding Gini index in cases where the estimate is country-specific to represent the level of inequality within each country. The Gini coefficients were obtained from World Bank Data. We also established dummy variables to determine whether the country is in Asia, South America, North America, or Europe. We also added the variable Study size, which counts the number of estimates produced by each study. This variable enables examination of potential skewness in studies with fewer reported estimates. Theoretically, studies with only one or two final estimates may have a higher inclination to manipulate results compared to studies with multiple specifications.

Finally, we created two variables to account for the prevalence of estimates derived from a common test type. The first variable indicates data obtained directly from the OECD studies, which employ similar methodologies. The second variable encompasses all studies that use PISA results as an indicator of achievement, whether or not they conduct their own estimations. The latter variable represents a subset of the former, as it includes studies using PISA results in addition to those directly published by the OECD.

Estimation approach

To delve deeper into the heterogeneity, we categorized the estimation methods used by primary studies, including Ordinary Least Squares, Fixed Effects, and Hierarchical Linear Modelling. For less frequent estimation approaches, we used the variable Other method. Additionally, we encoded a binary variable to indicate whether attempts were made to mitigate the omitted variable bias. We also encoded whether prior achievement was used as a control variable. This practice has been shown to have a significant downward impact on the strength of the effect under consideration (Marks 2015). Finally, we captured information on whether the model included multiple average SES variables, as such a practice may lead to smaller effect sizes.

Publication characteristics

We included the number of citations for each paper as an indicator of its scholarly impact. Papers with higher citation counts are often associated with the use of rigorous statistical procedures and publication in reputable journals. Additionally, we encoded the publication years, spanning from 1986 to 2023. Over the years, advancements in research methodologies may have led to more precise findings. Moreover, the dynamics of peer effects on socioeconomic status could have evolved due to improvements in equality or the implementation of policies addressing this issue. We also calculated the number of citations per year, which provides a more standardized measure. This allows us to assess the relative importance of each paper within the context of its publication time frame.

We further encoded whether the publication status of the study, as being accepted in a journal, indicates that peer review and validation were conducted. This step aims to ensure that appropriate methods are used. Moreover, we categorized the discipline from which each paper originates, encompassing economics, social science, and psychology. Finally, we encoded whether the primary focus of the study was on peer effects. If peer effects were the central concern, it is likely that researchers would devote more attention to employing appropriate methodologies in their investigations.

Table 5.1: Description and summary statistics of encoded variables

Variable	Description	Mean	SD
Effect	The effect of a one standard deviation increase in peer socioeconomic status on academic achievement measured in standard deviations.	0.333	0.276
Standard Error	The standard error of the effect.	0.054	0.003
Socioeconomic status measurement			
SES SD	Standard deviation of SES variable.	1.608	4.783
Avg SES Class Level	at which level is average SES measured, =1 if class, otherwise cohort or school.	0.127	0.333
SES Home Resources	=1 if SES variable reflects home resources.	0.543	0.499
SES Parent. Income	=1 if SES variable reflects home resources.	0.0223	0.148
SES Parent. Education	=1 if SES variable includes parental education	0.71	0.454
SES Parent. Occupation	=1 if SES variable captures occupation of parents.	0.8	0.401
SES Dichotomous	=1 if the measure is dichotmous	0.06	0.238
SES Composite	=1 if SES is measured as composite of home resources, education and occupation of parents	0.661	0.474
SES Combined	=1 if SES variable captures more aspects at the same time i.e. mother education and father education)	0.795	0.404
Academic achievement test type			
SD of AA variable	Standard deviation of AA variable	80.71	37.092
AA Test: Language	=1 type of achievement test is language	0.448	0.498
AA Test: Math	=1 type of achievement test is math.	0.37	0.483
AA Test: Science	=1 type of achievement test is science.	0.154	0.361
AA Test: General	=1 type of test is General Academic Achievement.	0.027	0.161
Sample characteristics			
Study Size	Number of estimates encoded from the study.	37.521	22.018
Sample Size	Number of observations used to estimate the effect	34116.281	73840.426
Average Student Age	average age of the student in the sample	14.317	1.51
Country-level Data	=1 if data is agregated on a country level	0.931	0.254
Country	Name of the observed country		
Country: Asia	=1 if country located in Asia	0.127	0.333
Country: South America	=1 if country located in South America	0.207	0.406
Coutry: North America	=1 if country located in North America	0.078	0.268
Country: Europe	=1 if country located in Europe	0.439	0.497
Gini coefficient	Gini coefficient for each country (World Bank Data)	35.444	6.499
PISA data	=1 if study uses PISA data as source	0.704	0.457
OECD	=1 if study was published by OECD	0.673	0.47
Estimation approach			
Method: OLS	=1 if the authors use Ordinary least squares.	0.149	0.357
Method: FE	=1 if the authors use Fixed-effects estimation.	0.136	0.343
Method: HLM	=1study uses Hierarchical liner modelling	0.715	0.452
Method: Other	1= if study uses CML or B2SLS	0.047	0.211
Overcome OVB	=1 if study attempts to overcome omitted variable bias	0.154	0.361
Prior attainment control	=1 if a study includes prior attainment as a covariate	0.065	0.246
Average SES > 1	= 1 if more than one average SES-variable is present in one model	0.122	0.328
Publication characteristics			
Publication Year	Year when the study was published	2007.873	6.004
Ctations per year	Number of citations per year.	18.031	14.195
Citations	Number of Google Scholar citations of May 2023.	262.702	252.676
Published	=1 if the study was published in a journal.	0.463	0.499
Discipline: Economics	=1 if study comes from field of Economics.	0.176	0.381
Discipline: Soc. Science	=1 if study comes from field of Social Science.	0.815	0.389
Discipline: Psychology	=1 if study comes from field of Psychology.	0.007	0.082
Primary concern	=1 if study of peer effects is a primary concern	0.176	0.381

Note: Table provides summary statistics and description of encoded variables. SD = standard deviation.

5.2 Model averaging

Once we have clarified the final structure of the data set, our attention turns towards conducting rigorous tests to explore heterogeneity in the data set. Ideally, we would conduct a regression analysis wherein we regress the collected estimates of peer effect on all of the variables specified in table 5.1. However, given the large number of regressors involved, there is a notable probability that a significant portion of these variables will be redundant. This redundancy, in turn, has the potential to compromise the precision of parameter estimates for the more significant variables in the regression model. In other words, we face substantial model uncertainty, which needs to be addressed (Eicher *et al.* 2011). Model uncertainty is a common problem in meta-analyses, as it is caused by the complexity of the literature and the consequent large number of aspects in which individual studies and estimates differ (Steel 2020). To tackle this issue, we employ Bayesian and frequentist model averaging techniques. The Bayesian model averaging method estimates the likelihood of including each individual explanatory variable in the underlying model. On the other hand, the frequentist approach, although more computationally demanding, does not necessitate the selection of prior probabilities and provides a robustness check.

We now delve deeper into the BMA. This approach involves running regressions including all possible combinations of the explanatory variables. We simplify this process using the Metropolis-Hastings algorithm from the *bms* package in R developed by Zeugner & Feldkircher (2015). This algorithm focuses on exploring the most probable models, reducing computational complexity (Cazachevici *et al.* 2020). The posterior model probabilities serve as an indicator of the likelihood of each model. Subsequently, the posterior means are calculated by weighting the estimated coefficients across all models, taking into account their respective posterior model probabilities. In a similar vein, the posterior inclusion probability (PIP) of a variable is defined as the sum of the posterior model probabilities for all models that include the respective variable. It can be considered analogous to statistical significance (Steel 2020). In essence, the PIP reflects the importance of a variable in the average model and indicates the likelihood of its inclusion in the final model. Thus, a higher PIP value corresponds to a greater level of importance. To interpret the significance of each variable based on its PIP, we refer to the guidelines proposed by Kass & Raftery (1995). They propose PIP values between 0.5 and 0.75 to indicate weak evidence of an effect, values between 0.75 and 0.9 to indicate a positive effect, values between 0.9 and 0.99 to indicate a strong effect, and values greater than 0.99 to indicate a decisive effect. We highlighted variables with a PIP of at least 0.5. For further information on BMA, please see Raftery *et al.* (1997) and Eicher *et al.* (2011).

It is noteworthy that we deviate from the conventional approach and adopt a dilution prior instead of a uniform model prior. This choice is motivated by the need to address potential collinearity issues within the model. The dilution prior, as proposed George *et al.* (2010), primarily serves the purpose of mitigating collinearity by incorporating the model probabilities and the determinant of the correlation matrix of the independent variables. This approach assigns greater weights to variables with lower correlations. We opt for the dilution as we use numerous similar variables in our model, which increases the likelihood of collinearity. Moreover, the large number of variables used in our analysis may exacerbate the aforementioned issues. To observe the correlation between the variables, see Figure B.1 in Appendix B.

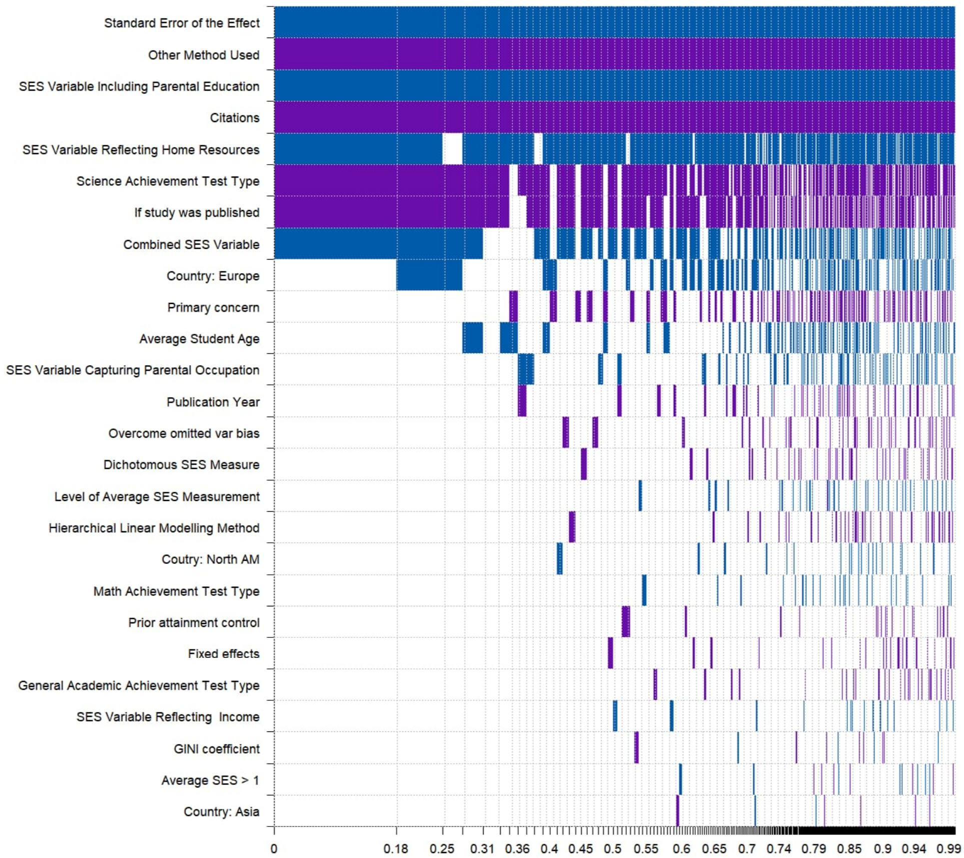
We also provide a robustness check using Frequentist Model Averaging. Following the approach of Gechert *et al.* (2022), we utilize Mallow's criteria as weights (Hansen 2007) and incorporate orthogonalization of the covariate space (Amini & Parmeter 2012). We adopt this approach as previously used method to reduce computational complexity cannot be used in this context. By employing this alternative method of model averaging, we can compare the results with our findings from BMA and further elucidate existing heterogeneity.

5.2.1 Overall results

Table 5.2 provides the numerical results of BMA and a robustness check using Frequentist Model Averaging (FMA). The graphical representation of the model averaging results can be found in Figure 5.1. Each column of the figure denotes a distinct individual regression model. The width of each column is proportional to the corresponding posterior model probability, reflecting the relative weight assigned to the model. The columns are arranged in descending order according to the posterior model probabilities. Conversely, each row within the figure represents a specific regression variable. The ordering of the rows is determined by the posterior inclusion probability, with variables having higher probabilities placed at the top in descending order. The cells are visually distinguished by their respective colors. Blue indicates a positive effect of the variable on the estimate of the peer effect, and violet represents a negative effect. The cell is left blank when a variable is not included in the model.

The implications drawn from the figure suggest that approximately two-fifths of the explanatory variables employed exhibit some degree of usefulness in explaining the heterogeneity observed in the reported estimates of the peer effect of SES. Furthermore, the coefficient signs for these variables remain consistent and robust across almost all of the models considered.

Figure 5.1: Model inclusion in Bayesian model averaging



Notes: The figure presents the results of Bayesian model averaging using the uniform g-prior suggested by Eicher *et al.* (2011) and the dilution model prior recommended by George *et al.* (2010). Each row represents one explanatory variable, ordered based on their posterior inclusion probability. The columns correspond to individual models. A blue color indicates that the variable positively affects the outcome, and violet suggests a negative impact. White cells signify that the variable was not included in the model. The models are ordered by their posterior model probabilities, with the most preferred models displayed on the left. For numerical results, see Table 5.2. For overview of the variables, see Table 5.1.

Table 5.2: Model averaging numerical results

Regressand:	Bayesian model averaging			Frequentist model averaging		
Effect of peer SES on AA	Post. mean	Post. SD	PIP	Coef.	SE	p-value
Intercept	8.607	NA	**1.000**	53.024	67.522	0.432
Standard Error	1.576	0.315	**1.000**	1.320	0.333	**0.000**
Socioeconomic status measurement						
SES Class Level	0.004	0.024	0.050	0.124	0.062	**0.047**
SES: Home Resources	0.103	0.048	**0.911**	0.106	0.046	**0.022**
SES Parental Income	0.002	0.017	0.027	0.072	0.079	0.361
SES: Parental Education	0.140	0.031	**1.000**	0.174	0.035	**0.000**
SES: Parental Occupation	0.012	0.036	0.121	0.035	0.063	0.573
SES Dichotomous	-0.006	0.031	0.060	-0.154	0.084	0.069
SES Combined	0.067	0.056	**0.652**	0.073	0.042	0.085
Academic achievement test type						
AA: Math	0.001	0.010	0.036	0.012	0.030	0.683
AA: Science	-0.095	0.051	**0.842**	-0.095	0.036	**0.008**
AA: GAA	-0.003	0.021	0.031	-0.162	0.084	0.054
Sample characteristics						
Average Student Age	0.049	0.113	0.193	0.115	0.148	0.439
Country: Asia	0.000	0.003	0.012	0.036	0.032	0.267
Coutry: North AM	0.002	0.012	0.037	0.063	0.043	0.144
Country: Europe	0.019	0.029	0.335	0.060	0.028	**0.028**
Gini coefficient	-0.001	0.012	0.020	0.017	0.072	0.818
Estimation approach						
Method: Other	-0.708	0.087	**1.000**	-0.575	0.109	**0.000**
Method: HLM	-0.002	0.013	0.049	-0.060	0.042	0.154
Method: FE	-0.001	0.010	0.031	-0.005	0.041	0.910
Prior attainment control	-0.002	0.014	0.033	-0.023	0.051	0.653
Average SES var > 1	0.000	0.008	0.020	0.009	0.066	0.894
Overcome OVB	-0.004	0.018	0.065	-0.058	0.047	0.217
Publication characteristics						
Citations	-0.075	0.020	**0.999**	-0.085	0.024	**0.000**
Published	-0.092	0.055	**0.791**	-0.038	0.052	0.457
Publication Year	-1.083	4.065	0.094	-6.952	8.857	0.433
Primary concern	-0.020	0.045	0.203	-0.129	0.056	**0.022**

Note: The table displays the results of Bayesian and Frequentist model averaging. Post. mean = Posterior Mean, Post. SD = posterior standard deviation, PIP = Posterior Inclusion Probability, Coef. = Coefficient, SE = Standard Error. Variables with a PIP value greater than 0.5 or a p-value below 0.05 are highlighted, indicating a higher probability of inclusion. See Table 5.1 for a detailed explanation of the variables.

We proceed to elaborate on the numerical results of model averaging from Table 5.2. The BMA method identified eight significant variables, as indicated by their Posterior Inclusion Probability (PIP) values exceeding 0.5. The first significant finding pertains to publication bias, represented by the large (1.576) positive effect assigned to standard error. Despite explicit control for different aspects of study design, a consistently positive relationship between estimates and standard errors is observed. Notably, Bayesian model averaging assigns a posterior inclusion probability of 1 to the standard error, while frequentist model averaging yields a p-value below 0.0001, confirming its statistical significance. Based on this observation, it can be confidently stated that the correlation between estimates and their standard errors is not simply the result of an omitted variable bias. BMA and FMA, therefore, serve as robustness checks for the results presented in the previous chapter on Publication bias.

Our focus now shifts to significant variables within the SES measurement category. With three variables obtaining a PIP above 0.5, it becomes evident that the SES measurement is essential. The associated low p-values from FMA further reinforce this notion. Specifically, the variable representing SES measurement incorporating parental education received a PIP of 1. Similarly, the SES measurement variable involving home resources obtained a PIP of 0.91. Both of these variables obtained positive coefficients. Lastly, it is worth noting that the variable indicating the combination of multiple indicators to derive socioeconomic status (SES) appears to have a noticeable positive impact. This finding aligns with the notion that integrating various inputs allows a more accurate representation of SES. It is worth noting that this variable was not found significant in the FMA.

The choice of academic achievement test also appears to have an impact, as evidenced by the variable representing the Science test type, which obtained a Posterior Inclusion Probability (PIP) of 0.842 and a coefficient of -0.095. Notably, the reference variable in this context is the Language test.

Furthermore, the variable representing the use of the Other methods received a PIP of 1. It is important to note that the reference variable within this category is the OLS method, while all the estimated coefficients for the methods category are negative. Therefore, this finding provides more insights into the effect of the OLS method itself rather than the "other method" category, which indicates various methods that were used only a few times. The results suggest that studies employing the Ordinary Least Squares (OLS) method consistently yield higher estimates. This observation is sensible, considering that

OLS serves as a baseline method and may be susceptible to bias. In more recent studies, researchers tend to employ more rigorous methods, such as Hierarchical Modeling or Fixed Effects, to mitigate potential biases.

Finally, we touch upon significant variables from the Publication characteristics category. The variable representing the number of citations obtained a PIP of 0.999, and the variable indicating that the study was published obtained a PIP of 0.791. It is noteworthy that both variables are estimated to have a negative sign, which is interesting. The negative coefficient of the variable representing publishing can be explained by the fact that OECD studies, on average, yield larger estimates but are not considered published. Also, the signs of both variables could be explained by the peer-review process required for publication. This process often leads to more rigorous methods and may result in smaller effect estimates. This idea aligns with the notion that the true effect size in the literature might be smaller on average, which is consistent with findings from the previous chapter.

Table 5.2 additionally presents a robustness check through the application of Frequentist Model Averaging (FMA). For the sake of clarity, we have highlighted variables in which the associated p-value is below 0.05. The results align closely with the Bayesian Model Averaging (BMA) approach, as both approaches indicate similar directions and significance for most of the variables. However, there are a few exceptions to note. In contrast to BMA, the FMA approach found the constant term, Combined SES, and Publication status to be statistically insignificant. Conversely, in addition to the variables deemed significant by the BMA approach, FMA identified the significance of three additional variables: SES Class Level, Europe and Primary concern.

The variable representing the average socioeconomic status (SES) measurement at the class level obtained a large positive coefficient, which aligns with intuition. This result suggests that the proximity of peers within the same class tends to have a stronger influence on students compared to those from the entire school or cohort. Additionally, FMA assigned a positive coefficient to the variable indicating the sample taken in Europe. Notably, South America serves as the reference category for this group of variables. Finally, we highlight the significant negative coefficient assigned to the primary concern variable. This can be explained by researchers' emphasis on addressing the inherent biases associated with the analysis of peer effects, resulting in substantially lower estimates.

To conclude, our findings exhibit a degree of similarity to the results of a previous meta-analysis by Van Ewijk & Sleegers (2010) that employed a fixed effects meta-regression approach. While both studies observed mostly coefficients with similar signs, there were differences in the identification of significant variables. These disparities can be attributed to three main factors. Firstly, the use of different estimation approaches. Secondly, our study incorporated a larger number of studies, potentially providing a more comprehensive analysis. Lastly, our study also included a broader range of variables.

5.2.2 Results excluding OECD studies

In order to address the notable effect of OECD (2001; 2003; 2004; 2005; 2007; 2010; 2014) studies on the results of previous BMA, we chose to conduct a separate analysis excluding these studies. We used an identical setup as in the previous model averaging.

Figure 5.2 presents the visual outcomes of model averaging. We initially hypothesized that the similarity in OECD estimates would lead to an inflation of certain important effects. However, the results do not strongly support this hypothesis, as can be seen in the graphical results. The overall explanatory power is diminished, and fewer variables were found to be statistically significant. We did, however, identify a number of significant variables that differ from our previous findings.

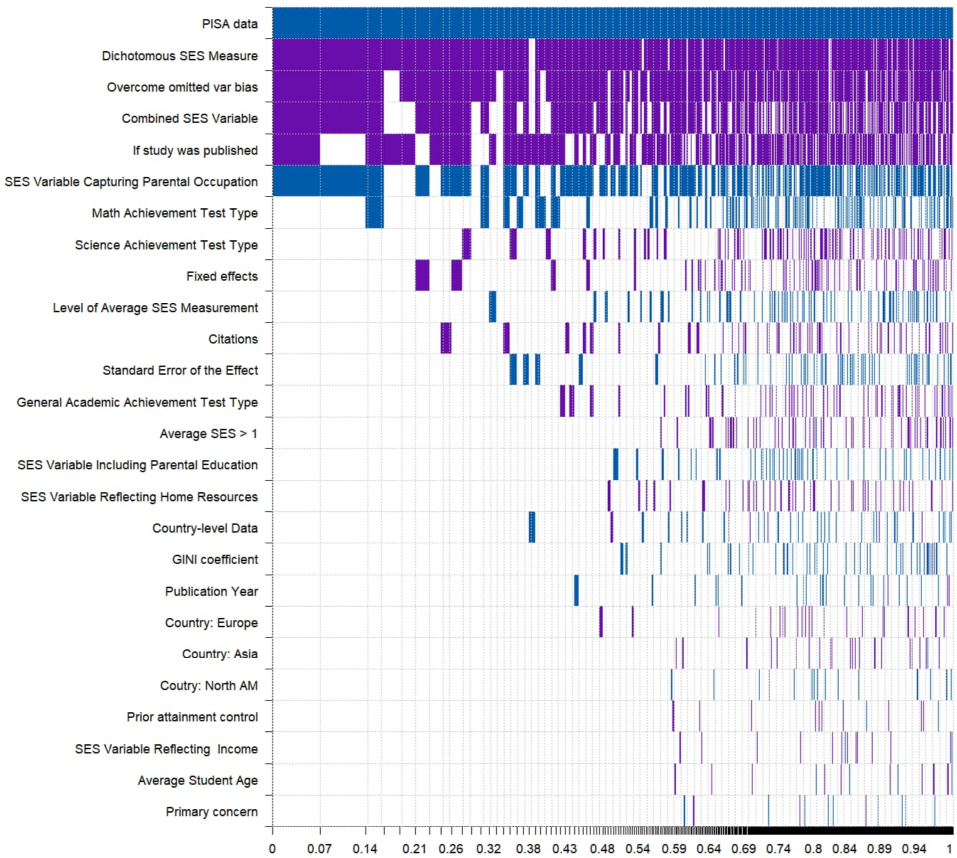
Table 5.3 presents six variables with a PIP over 0.5, sorted based on their PIP values. The decisive effect was attributed to the variable indicating that the data used came from PISA tests. This means that even though we do not directly include studies from OECD (which use PISA tests), our conclusions remain affected by the results of PISA tests. The second most influential variable was the use of dichotomous SES measures, followed by an attempt to address omitted variable bias. Additionally, the variables representing combined SES measures and publication status were identified as significant, consistent with the previous analysis. Lastly, using parental occupation as a measure of SES was found to be weakly significant. However, the parallel Frequentist model averaging results did not provide support for all of these findings. Specifically, this analysis indicated that the last two variables with the lowest PIP were deemed insignificant. The correlation matrix (Figure B.3) and Posterior model size and convergence (Figure B.4) are provided in Appendix B.

Table 5.3: Model averaging results - excluding OECD studies

Regressand:	Bayesian model averaging			Frequentist model averaging		
Effect of peer SES on AA	Post. mean	Post. SD	PIP	Coef.	SE	p-value
Intercept	-0.508	NA	1.000	30.803	53.382	0.564
PISA data	0.230	0.051	0.999	0.240	0.069	0.000
SES Dichotomous	−0.180	0.058	0.962	−0.161	0.071	0.023
Overcome OVB	−0.112	0.052	0.886	−0.107	0.040	0.008
SES Combined	−0.106	0.077	0.716	−0.124	0.055	0.024
Published	−0.137	0.108	0.700	−0.197	0.102	0.053
SES: Parental Occupation	0.075	0.063	0.646	0.078	0.059	0.185

Note: The table displays the results of Bayesian and Frequentist model averaging. Post. mean = Posterior Mean, Post. SD = posterior standard deviation, PIP = Posterior Inclusion Probability, Coef. = Coefficient, SE = Standard Error. Variables were ordered according to their PIP. See Table 5.1 for a detailed explanation of the variables.

Figure 5.2: Graphical results of the alternative BMA



Notes: The figure presents the results of Bayesian model averaging excluding OECD studies. Each row represents one explanatory variable, ordered based on their posterior inclusion probability. The columns correspond to individual models. A blue color indicates that the variable positively affects the outcome, and violet suggests a negative impact. White cells signify that the variable was not included in the model. The models are ordered by their posterior model probabilities, with the most preferred models displayed on the left. Numerical results are provided in Table 5.3 and an overview of the variables is in Table 5.1.

Chapter 6

Final remarks

6.1 The Best-practise estimate

As proposed in guidelines by Havránek *et al.* (2020), we now proceed to estimate the best practice estimate using the Bayesian model averaging (BMA) approach introduced in the previous chapter. This method entails using the coefficients from the BMA model and incorporating the characteristics that represent an ideal scenario for the effect measurement. However, it is important to emphasize that this procedure is inherently subjective and should primarily be considered as an additional robustness check rather than a means of presenting novel findings. Then, we compare our model to four primary studies, in which we input the values of the selected variables into the BMA model to derive estimates from "ideally designed" studies.

When modeling the best practice estimate for many variables, the optimal value is not clear. In such cases, we employed sample average as a default value. Additionally, we subjectively selected several characteristics that define an "ideal study," which we now elaborate on. First, we set the standard error to 0 to approximately correct for publication. Further, we set the variable indicating Dichotomous measure of SES to 0, as this approach is undesirable. We also set the study as published, as peer review leads to more appropriate methods. Subsequently, we prefer peer effects to be the primary concern, as it is commonly observed that authors who center their focus on peer effects are more likely to account for potential biases. Then, we prefer the SES variable to be Combined, as more inputs lead to a more precise approximation of SES. Lastly, we ensured that the control variable for prior achievement was included in the model. By doing so, we aimed to mitigate the risk of the estimated effect serving as a proxy for individual abilities. The results can be seen in Table 6.1.

Our best-practice estimate is 0.195, with a confidence interval of (0.113; 0.277). This result is considerably lower than the sample average of 0.333. However, this estimate aligns with the results of multiple tests conducted to detect publication bias. We now explain our choice of studies. First, we selected the publication with the highest number of citations, both absolutely and per year, which is study Sui-Chu & Willms (1996). Then we selected a study by Berkowitz (2022), as it is one of the latest in our data set. Furthermore, the author has extensively explored the topic and has already produced multiple studies on the subject matter. Also, we used the study by Schneeweis & Winter-Ebmer (2007), which is both intensively cited and, in our opinion, well-designed. Finally, we decided to also provide an estimate for the OECD study, as its estimates represent a significant portion of our data set. Notably, some of the resulting estimates are accompanied by noticeably wide confidence intervals.

Table 6.1: Best-practice estimate

Author	**Best-practise estimate**	**95% CI**
Author	0.195	(0.113; 0.277)
Sui-Chu & Willms (1996)	0.055	(-0.039; 0.149)
Berkowitz (2022)	0.346	(0.193; 0.499)
Schneeweis & Winter-Ebmer (2007)	0.188	(0.072; 0.304)
OECD (2005)	0.196	(0.141; 0.251)

Notes: The table contains best-practice estimates according to the author and four selected studies. CI = confidence interval. The 95% confidence interval bounds are approximated using OLS with clustered standard errors at the study level.

6.2 Economic significance

As a final step, we investigate the economic significance of the eight variables that obtained a posterior inclusion probability of at least 0.5 in BMA. Table 6.2 shows the effect of the change of each variable on the peer effect, controlling for all other factors. We report the effects of a one standard deviation increase along with the transition from minimum to maximum value. When studying dummy variables, focusing on the transition spanning from the minimum to the maximum value is often more meaningful. Conversely, a one-standard-deviation change usually provides more valuable information for continuous variables. Additionally, we express these effects as percentage changes relative to our recently derived subjective best-practice estimate.

When considering the variables in terms of the effect resulting from a one standard deviation increase, the most substantial impact is observed when employing Other methods (it is important to note that other denotes more advanced methods, whereas the reference variable for the group was the Ordinary Least Squares (OLS) method). The number of citations stands out as the second most influential factor. Similarly, the standard error carries significant importance, serving as a proxy for publication bias. Furthermore, the inclusion of parental education in the measurement of SES exhibits a similarly strong effect. Likewise, the order of variable impacts remains consistent when considering the perspective of transitioning from the minimum to maximum values.

Table 6.2: Economic significance of key variables

	One SD increase		Maximum change	
Variable	Effect	% of BPE	Effect	% of BPE
Standard Error	0.069	35.15%	0.360	184.81%
Citations	-0.088	-45.28%	-0.575	-294.95%
Published	-0.046	-23.47%	-0.092	-347.01%
Method: Other	-0.150	-76.74%	-0.708	-363.05%
AA: Science Test	-0.034	-17.50%	-0.095	-48.47%
SES: Home Resources	0.052	26.44%	0.103	53.02%
SES: Parental Education	0.064	32.58%	0.140	71.76%
SES: Combined	0.027	13.94%	0.067	34.51%

Notes: The table shows the ceteris paribus changes in the reported effects of peer socioeconomic status on academic achievement implied by changes in the several relevant variables. Only variables with a Bayesian Model Averaging (BMA) Posterior Inclusion Probability (PIP) exceeding 0.5 are included in the analysis. One SD increase refers to the effect of one standard deviation increase of a specific variable. Additionally, the maximum change indicates the variation in the effect when the variable is increased from its minimum to its maximum value. The reference best-practice estimate is 0.195. BPE = Best-Practice estimate, See table 5.1 for an overview of the variables.

Chapter 7

Conclusion

This thesis delves deep into the effect of peer socioeconomic status on academic achievement, a topic widely discussed across various fields. We employ state-of-the-art methods to address publication bias and model uncertainty. Our data set includes 449 estimates from 40 studies, which were recalculated for comparability. Specifically, we examine the effect of a one standard deviation change in average socioeconomic status (SES) on standard deviations of academic achievement.

The topic of peer effects in education holds significance in various academic debates, such as the school choice and ability grouping debates, as well as the broader discourse on optimizing educational systems (Rangvid 2003). Furthermore, obtaining accurate information on the true effect size has practical implications, as it can be useful for policymakers, school principals, and parents (Paloyo 2020). Nevertheless, the literature on the effect provides mixed results, and previous meta-analysis by Van Ewijk & Sleegers (2010) did not rigorously assess publication bias.

We address the research gap by examining the publication bias employing a battery of 18 various statistical tests, obtaining an effect size ranging from 0.117 to 0.338. Furthermore, the majority of the tests conducted indicate the presence of publication bias in the literature while achieving statistical significance at the 5% level. These results contrast with a prior meta-analysis by Van Ewijk & Sleegers (2010) that did not detect a significant presence of publication bias.

Moreover, we collected over 40 different variables capturing the characteristics of studies to explain the heterogeneity in our data set. The encoded variables contain information on the measurement of SES, measurement of academic achievement, sample characteristics, estimation approach, and publication characteristics. To explain the heterogeneity, we use Bayesian model averaging (BMA) and frequentist model averaging (FMA). The application of BMA reveals a statistically significant positive relationship between the effect size and the following variables: Standard error, the SES variable reflects parental education, the SES variable reflects home resources, and the SES variable is a composite of more inputs. We obtain a significant negative relationship for the following variables: Use of other methods (i.e., more advanced, having the OLS method as a reference variable), number of citations, the study being published, and Academic achievement test type being science. On top of this, the FMA also indicated the significance of the level at which SES is measured (positive effect), whether the data is from Europe (positive effect), and if the study of peer effects is a primary concern of the paper (negative effect). Our findings regarding the drivers of the effect are mostly consistent with the findings of Van Ewijk & Sleegers (2010).

To address the influence of 7 OECD studies on the previous model averaging results, we perform a separate analysis by excluding these studies. The new analysis reveals additional significant variables; however, the overall explanatory power is reduced. In addition to the variables identified in the previous model, the use of PISA data, a dichotomous measure of socioeconomic status (SES), an attempt to overcome omitted variable bias, and parental occupation as a measure of SES emerged as significant variables. However, not all of these findings are supported by the parallel frequentist model averaging results.

Finally, we present a subjective best-practice estimate and compare it with four primary studies from our data set, providing a contextual comprehension of the findings. We identify consistent patterns across different specifications, supported by an analysis of the economic significance of key variables from BMA. Similarly to previous findings, we observe a noticeable influence of publication bias. Notably, our proposed best-practice estimate of 0.2 is lower than the robustness check of 0.31 provided by Van Ewijk & Sleegers (2010).

Lastly, we acknowledge the limitations of this thesis and put forth suggestions for future extensions. Firstly, as we have shown, segregation and the resulting effect vary significantly across countries. Consequently, drawing definitive conclusions regarding the true underlying effect becomes challenging. Secondly, greater attention could have been given to the OECD studies, which may introduce a certain degree of upward bias to the resulting estimates of the true effect. Regarding potential extensions of this work, it may be interesting to collect estimates from studies that use methods that cannot be recalculated into standardized effects but instead employ partial correlation coefficients. Also, the inclusion of a larger number of studies could have enhanced the scope of the analysis, enabling it to examine potential gender differences and determine whether the effect diminishes with increasing ability. Additionally, we propose collecting more variables, such as information on whether hierarchical linear modeling was conducted on 2 or 3 levels and information on control variables in the model specification. The motivation for gathering more information on control variables stems from the studies by Marks (2015; 2016). Their findings suggest that when accounting for factors such as prior attainment, parental involvement, and intellect, the observed effect becomes statistically insignificant.

Bibliography

van Aert, R. C. & M. Van Assen (2021): "Correcting for publication bias in a meta-analysis with the p-uniform* method." *Manuscript submitted for publication Retrieved from: https://osfio/preprints/bitss/zqjr92018* .

Aguinis, H., D. R. Dalton, F. A. Bosco, C. A. Pierce, & C. M. Dalton (2011): "Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact." *Journal of Management* **37(1)**: pp. 5–38.

Allenby, G. M. & P. E. Rossi (2006): "Hierarchical bayes models." *The handbook of marketing research: Uses, misuses, and future advances* pp. 418–440.

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

Ammermueller, A. & J.-S. Pischke (2009): "Peer effects in european primary schools: Evidence from the progress in international reading literacy study." *Journal of Labor Economics* **27(3)**: pp. 315–348.

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

Aram, D. & I. Levin (2001): "Mother–child joint writing in low ses: Sociocultural factors, maternal mediation, and emergent literacy." *Cognitive Development* **16(3)**: pp. 831–852.

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

Bajzik, J., T. Havranek, Z. Irsova, & J. Schwarz (2020): "Estimating the armington elasticity: The importance of study design and publication bias." *Journal of International Economics* **127**: p. 103383.

Bankston III, C. & S. J. Caldas (1996): "Majority african american schools and social injustice: The influence of de facto segregation on academic achievement." *Social Forces* **75(2)**: pp. 535–555.

Bankston III, C. L. & S. J. Caldas (1998): "Family structure, schoolmates, and racial inequalities in school achievement." *Journal of Marriage and the Family* pp. 715–723.

Berkowitz, R. (2021): "School climate and the socioeconomic literacy achievement gap: Multilevel analysis of compensation, mediation, and moderation models." *Children and Youth Services Review* **130**: p. 106238.

Berkowitz, R. (2022): "School matters: The contribution of positive school climate to equal educational opportunities among ethnocultural minority students." *Youth & Society* **54(3)**: pp. 372–396.

Berkowitz, R., H. Glickman, R. Benbenishty, E. Ben-Artzi, T. Raz, N. Lipshtat, & R. A. Astor (2015): "Compensating, mediating, and moderating effects of school climate on academic achievement gaps in israel." *Teachers College Record* **117(7)**: pp. 1–34.

Betts, J. R. (1998): "The impact of educational standards on the level and distribution of earnings." *The American Economic Review* **88(1)**: pp. 266–275.

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

Borgatta, E. F. & D. J. Jackson (1980): *Aggregate data: Analysis and interpretation.* Sage Publications Beverly Hills/London.

Buckingham, J., K. Wheldall, & R. Beaman-Wheldall (2013): "Why poor children are more likely to become poor readers: The school years." *Australian Journal of Education* **57(3)**: pp. 190–213.

Cala, P., T. Havranek, Z. Irsova, J. Matousek, & J. Novak (2022): "Financial incentives and performance: A meta-analysis of economics evidence." .

Caldas, S. J. & C. Bankston III (1998): "The inequality of separation: Racial composition of schools and academic achievement." *Educational Administration Quarterly* **34(4)**: pp. 533–557.

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

Cazachevici, A., T. Havranek, & R. Horvath (2020): "Remittances and economic growth: A meta-analysis." *World Development* **134**: p. 105021.

Coleman, J. S. (1966): *Equality of Educational Opportunity [summary Report].*, volume 1. US Department of Health, Education, and Welfare, Office of Education.

Cowan, C. D., R. M. Hauser, R. A. Kominski, H. M. Levin, S. R. Lucas, S. L. Morgan, & C. Chapman (2012): "Improving the measurement of socioeconomic status for the national assessment of educational progress: A theoretical foundation." *National Center for Education Statistics* **2012**.

De Clercq, M., B. Galand, & M. Frenay (2017): "Transition from high school to university: a person-centered approach to academic achievement." *European journal of psychology of education* **32**: pp. 39–59.

De Fraine, B., J. Van Damme, G. Van Landeghem, M.-C. Opdenakker, & P. Onghena (2003): "The effect of schools and classes on language achievement." *British educational research journal* **29(6)**: pp. 841–859.

Duncan, O. D., D. L. Featherman, & B. Duncan (1972): "Socioeconomic background and achievement." Seminar Press.

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

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

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

Elminejad, A., T. Havránek, & Z. Havránková (2022): "People are less risk-averse than economists think." *Technical report*, IES Working Paper.

Erola, J., S. Jalonen, & H. Lehti (2016): "Parental education, class and income over early life course and children's achievement." *Research in Social Stratification and Mobility* **44**: pp. 33–43.

Evans, W. N., W. E. Oates, & R. M. Schwab (1992): "Measuring peer group effects: A study of teenage behavior." *Journal of Political Economy* **100(5)**: pp. 966–991.

Feldman, J. M. (1984): "Meta-analysis: Cumulating research findings across studies." *Academy of Management. The Academy of Management Review (pre-1986)* **9(000001)**: p. 165.

Furukawa, C. (2019): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Evidence, and a New Correction Method (March 29, 2019)* .

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

George, E. I. *et al.* (2010): "Dilution priors: Compensating for model space redundancy." *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown* **6**: pp. 158–165.

Gerber, A., N. Malhotra *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)**: pp. 313–326.

Gutiérrez, G. (2023): "Is it socioeconomic or academic? disentangling sources of peer effects on student achievement." *British Journal of Sociology of Education* **44(1)**: pp. 144–163.

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

Harel Ben Shahar, T. (2022): *Ability and Ability Grouping.*

Harker, R. & P. Tymms (2004): "The effects of student composition on school outcomes." *School effectiveness and school improvement* **15(2)**: pp. 177–199.

Hauser, R. M. (1994): "Measuring socioeconomic status in studies of child development." *Child development* **65(6)**: pp. 1541–1545.

Havranek, T., Z. Irsova, K. Janda, & D. Zilberman (2015): "Selective reporting and the social cost of carbon." *Energy Economics* **51**: pp. 394–406.

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2021): "Skilled and unskilled labor are less substitutable than commonly thought." .

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

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

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

Havranek, T. & A. Sokolova (2020): "Do consumers really follow a rule of thumb? three thousand estimates from 144 studies say "probably not"." *Review of Economic Dynamics* **35**: pp. 97–122.

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

Henderson, V., P. Mieszkowski, & Y. Sauvageau (1978): "Peer group effects and educational production functions." *Journal of Public Economics* **10(1)**: pp. 97–106.

Hill, M. S. & S. P. Jenkins (1999): *Poverty among British children: chronic or transitory?* Citeseer.

Hutchison, D. (2003): "The effect of group-level influences on pupils' progress in reading." *British educational research journal* **29(1)**: pp. 25–40.

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

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

Iršová, Z., T. Havránek *et al.* (2010): "Measuring bank efficiency: a meta-regression analysis." *Prague Economic Papers* **19(10)**: pp. 307–328.

Izaguirre, A. & L. Di Capua (2020): "Exploring peer effects in education in latin america and the caribbean." *Research in Economics* **74(1)**: pp. 73–86.

Jencks, C., S. E. Mayer *et al.* (1990): "The social consequences of growing up in a poor neighborhood." *Inner-city poverty in the United States* **111**: p. 186.

Kang, C. (2007): "Classroom peer effects and academic achievement: Quasi-randomization evidence from south korea." *Journal of Urban Economics* **61(3)**: pp. 458–495.

Kartianom, K. & O. Ndayizeye (2017): "What's wrong with the asian and african students' mathematics learning achievement? the multilevel pisa 2015 data analysis for indonesia, japan, and algeria." *Jurnal Riset Pendidikan Matematika* **4(2)**: pp. 200–210.

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

Lee, V. E. & A. S. Bryk (1989): "A multilevel model of the social distribution of high school achievement." *Sociology of education* pp. 172–192.

López, V., M. Salgado, & R. Berkowitz (2023): "The contributions of school and classroom climate to mathematics test scores: a three-level analysis." *School Effectiveness and School Improvement* **34(1)**: pp. 43–64.

Ma, X. & D. A. Klinger (2000): "Hierarchical linear modelling of student and school effects on academic achievement." *Canadian Journal of Education/Revue canadienne de l'education* pp. 41–55.

Marks, G. N. (2015): "Are school-ses effects statistical artefacts? evidence from longitudinal population data." *Oxford Review of Education* **41(1)**: pp. 122–144.

Marks, G. N. (2016): "The relative effects of socio-economic, demographic, non-cognitive and cognitive influences on student achievement in australia." *Learning and Individual Differences* **49**: pp. 1–10.

Maxwell, S., K. J. Reynolds, E. Lee, E. Subasic, & D. Bromhead (2017): "The impact of school climate and school identification on academic achievement: Multilevel modeling with student and teacher data." *Frontiers in psychology* **8**: p. 2069.

McEwan, P. J. (2003): "Peer effects on student achievement: Evidence from chile." *Economics of education review* **22(2)**: pp. 131–141.

McEwan, P. J. (2004): "The indigenous test score gap in bolivia and chile." *Economic development and cultural change* **53(1)**: pp. 157–190.

Mizala, A. & F. Torche (2012): "Bringing the schools back in: the stratification of educational achievement in the chilean voucher system." *International Journal of Educational Development* **32(1)**: pp. 132–144.

Mueller, C. W. & T. L. Parcel (1981): "Measures of socioeconomic status: Alternatives and recommendations." *Child development* pp. 13–30.

Neuman, S. B. & D. C. Celano (2015): *Giving our children a fighting chance: Poverty, literacy, and the development of information capital.* Teachers College Press.

OECD (2001): *Knowledge and skills for life: First results from the OECD programme for International Student Assessment (PISA).* Paris, France: OECD Publishing.

OECD (2003): *Literacy skills for the world tomorrow: Further results from PISA 2000.* Paris, France: OECD Publishing.

OECD (2004): *Learning for Tomorrow's World: First Results from PISA 2003.* Paris, France: OECD Publishing.

OECD (2005): *School factors related to quality and equity: Results from PISA 2000.* Paris, France: OECD Publishing.

OECD (2007): *PISA 2006: Science competencies for tomorrow's world.* Paris, France: OECD Publishing.

OECD (2010): *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science (volume I).* Paris, France: OECD Publishing.

OECD (2014): *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014).* OECD Publishing. Â© OECD 2014.

Opatrny, M., T. Havranek, Z. Irsova, & M. Scasny (2023): "Publication bias and model uncertainty in measuring the effect of class size on achievement." .

Opdenakker, M.-C., J. Van Damme, D. F. De Fraine, G. Van Landeghem, & P. Onghena (2002): "The effect of schools and classes on mathematics achievement." *School effectiveness and school improvement* **13(4)**: pp. 399–427.

Paloyo, A. R. (2020): "Peer effects in education: recent empirical evidence." In "The economics of education," pp. 291–305. Elsevier.

Paterson, L. (1991): "Socio-economic status and educational attainment: a multi-dimensional and multi-level study." *Evaluation & Research in Education* **5(3)**: pp. 97–121.

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

Rangvid, B. S. (2003): "Educational peer effects: quantile regression evidence from denmark with pisa 2000 data." *European Society for Population Economics* .

Rangvid, B. S. (2008): "School composition effects in denmark: quantile regression evidence from pisa 2000." *The Economics of Education and Training* pp. 179–208.

Rivkin, S. G. (2001): "Tiebout sorting, aggregation and the estimation of peer group effects." *Economics of Education Review* **20(3)**: pp. 201–209.

Robertson, D. & J. Symons (2003): "Do peer groups matter? peer group versus schooling effects on academic attainment." *Economica* **70(277)**: pp. 31–53.

Robinson, W. S. (2009): "Ecological correlations and the behavior of individuals." *International journal of epidemiology* **38(2)**: pp. 337–341.

RODRIGUEZ-HERNANDEZ, C. F., E. CASCALLAR, & E. KYNDT (2020): "Socioeconomic status and academic performance in higher education: A systematic review." *Educational Research Review* **29**: p. 100305.

SCHNEEWEIS, N. & R. WINTER-EBMER (2007): "Peer effects in austrian schools." *Empirical economics* **32**: pp. 387–409.

SIRIN, S. R. (2005): "Socioeconomic status and academic achievement: A meta-analytic review of research." *Review of educational research* **75(3)**: pp. 417–453.

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

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

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

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

STANLEY, T. D. & H. DOUCOULIAGOS (2017): "Neither fixed nor random: weighted least squares meta-regression." *Research synthesis methods* **8(1)**: pp. 19–42.

STANLEY, T. D., H. DOUCOULIAGOS, M. GILES, J. H. HECKEMEYER, R. J. JOHNSTON, P. LAROCHE, J. P. NELSON, M. PALDAM, J. POOT, G. PUGH *et al.* (2013): "Meta-analysis of economics research reporting guidelines." *Journal of economic surveys* **27(2)**: pp. 390–394.

STANLEY, T. D., S. B. JARRELL, & H. DOUCOULIAGOS (2010): "Could it be better to discard 90% of the data? a statistical paradox." *The American Statistician* **64(1)**: pp. 70–77.

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

Sui-Chu, E. H. & J. D. Willms (1996): "Effects of parental involvement on eighth-grade achievement." *Sociology of education* pp. 126–141.

Summers, A. A. & B. L. Wolfe (1977): "Do schools make a difference?" *The American Economic Review* **67(4)**: pp. 639–652.

Sun, L., K. D. Bradley, & K. Akers (2012): "A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: Pisa hong kong sample." *International Journal of Science Education* **34(14)**: pp. 2107–2125.

Thien, L. M. (2016): "Malaysian students' performance in mathematics literacy in pisa from gender and socioeconomic status perspectives." *The Asia-Pacific Education Researcher* **25(4)**: pp. 657–666.

Van Ewijk, R. & P. Sleegers (2010): "The effect of peer socioeconomic status on student achievement: A meta-analysis." *Educational research review* **5(2)**: pp. 134–150.

Webster, B. J. & D. L. Fisher (2000): "Accounting for variation in science and mathematics achievement: A multilevel analysis of australian data third international mathematics and science study (timss)." *School Effectiveness and School Improvement* **11(3)**: pp. 339–360.

White, K. R. (1982): "The relation between socioeconomic status and academic achievement." *Psychological bulletin* **91(3)**: p. 461.

Willms, J. D. (1986): "Social class segregation and its relationship to pupils' examination results in scotland." *American sociological review* pp. 224–241.

Wössmann, L. (2003): "Schooling resources, educational institutions and student performance: the international evidence." *Oxford bulletin of economics and statistics* **65(2)**: pp. 117–170.

York, T. T., C. Gibson, & S. Rankin (2015): "Defining and measuring academic success." *Practical assessment, research, and evaluation* **20(1)**: p. 5.

Young, D. J. & B. J. Fraser (1992): "School effectiveness and science achievement: Are there any sex differences?." .

Young, D. J. & B. J. Fraser (1993): "Socioeconomic and gender effects on science achievement: An australian perspective." *School Effectiveness and School Improvement* **4(4)**: pp. 265–289.

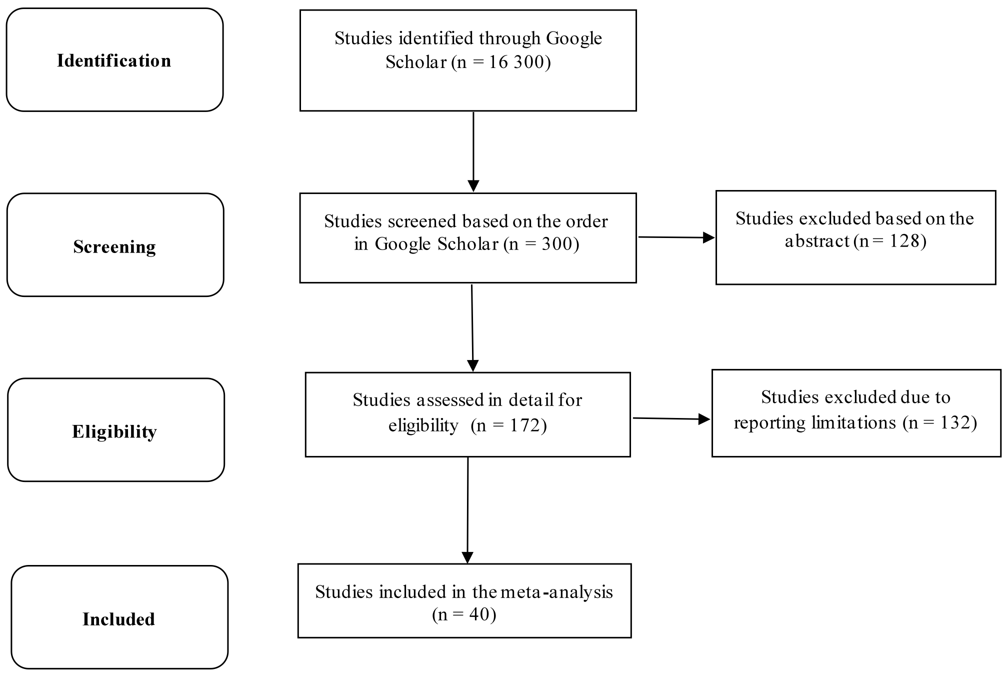
Zeugner, S. & M. Feldkircher (2015): "Bayesian model averaging employing fixed and flexible priors: The bms package for r." *Journal of Statistical Software* **68**: pp. 1–37.

Zimmer, R. W. & E. F. Toma (2000): "Peer effects in private and public schools across countries." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* **19(1)**: pp. 75–92.

Appendix A

Literature Search Details

Figure A.1: PRISMA flow diagram



Notes: Google Scholar's search functionality is comprehensive, allowing for full-text search. As a result, our search query is broad and comprehensive. We used the following query: "peer" OR "peer effect" OR "contextual effect" OR "peer influence" OR "composition" AND "socioeconomic influences" OR "socioeconomic status" OR "socioeconomic background" OR "classroom environment" AND "achievement". The search was conducted until May 1, 2023. PRISMA = The Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Additional information on PRISMA and meta-analysis reporting standards in general can be found in the paper from Havránek *et al.* (2020).

Table A.1: Studies included in the meta-analysis

Ammermueller & Pischke (2009)
Bankston III & Caldas (1996)
Bankston III & Caldas (1998)
Berkowitz *et al.* (2015)
Berkowitz (2021)
Berkowitz (2022)
Caldas & Bankston III (1998)
De Fraine *et al.* (2003)
Gutiérrez (2023)
Izaguirre & Di Capua (2020)
Kang (2007)
Kartianom & Ndayizeye (2017)
Lee & Bryk (1989)
López *et al.* (2023)
Ma & Klinger (2000)
Maxwell *et al.* (2017)
McEwan (2003)
McEwan (2004)
Mizala & Torche (2012)
OECD (2001)
OECD (2003)
OECD (2004)
OECD (2005)
OECD (2007)
OECD (2010)
OECD (2014)
Opdenakker *et al.* (2002)
Paterson (1991)
Rangvid (2003)
Rangvid (2008)
Rivkin (2001)
Sui-Chu & Willms (1996)
Schneeweis & Winter-Ebmer (2007)
Sun *et al.* (2012)
Thien (2016)
Webster & Fisher (2000)
Willms (1986)
Young & Fraser (1992)
Young & Fraser (1993)
Zimmer & Toma (2000)

Appendix B

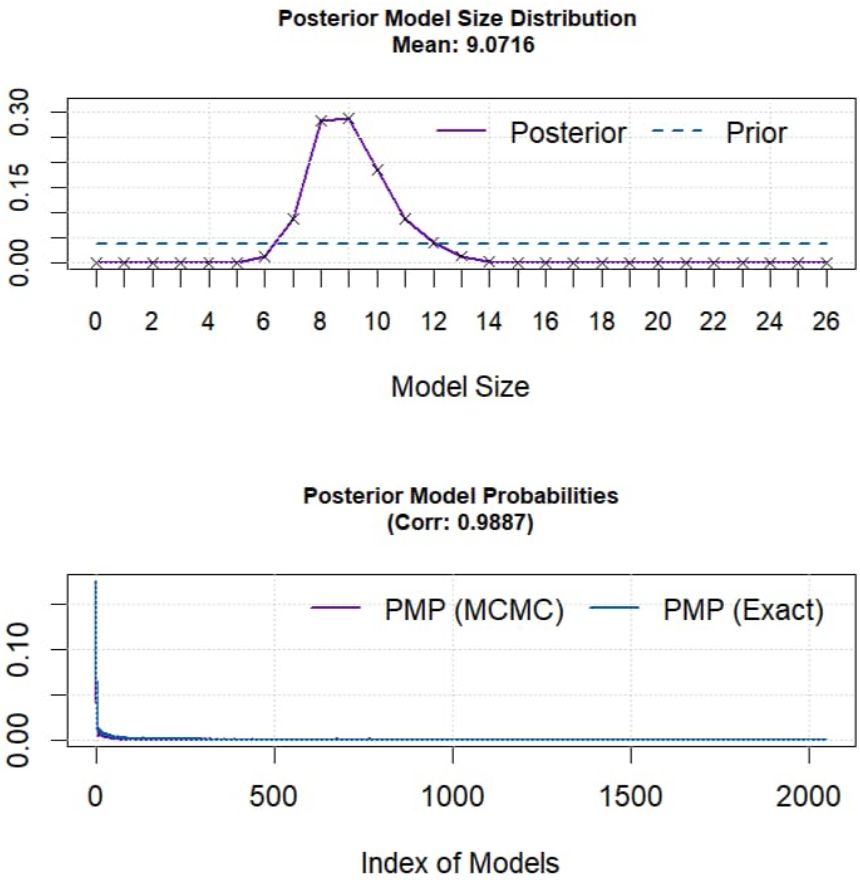
Additional information from BMA

B.1 Baseline BMA

Figure B.1: Correlation matrix of the explanatory variables

Notes: This figure depicts the correlation table for our baseline BMA. Blue color denotes positive correlation, while violet denotes negative correlation. For description of the variables see Table 5.1.

Figure B.2: Posterior model size and convergence of the BMA estimation

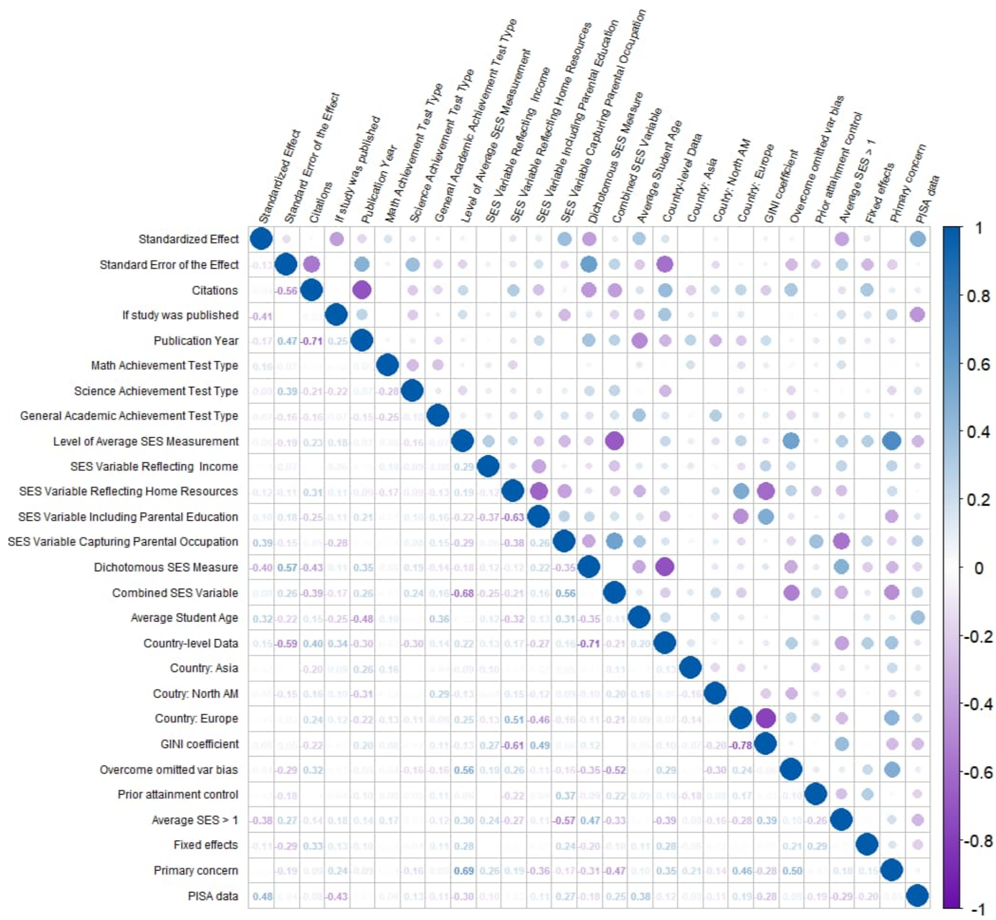


Notes: The figure presents the posterior model probabilities for different model sizes and posterior model size distribution of the baseline BMA.

B.2 Alternative BMA excluding OECD studies

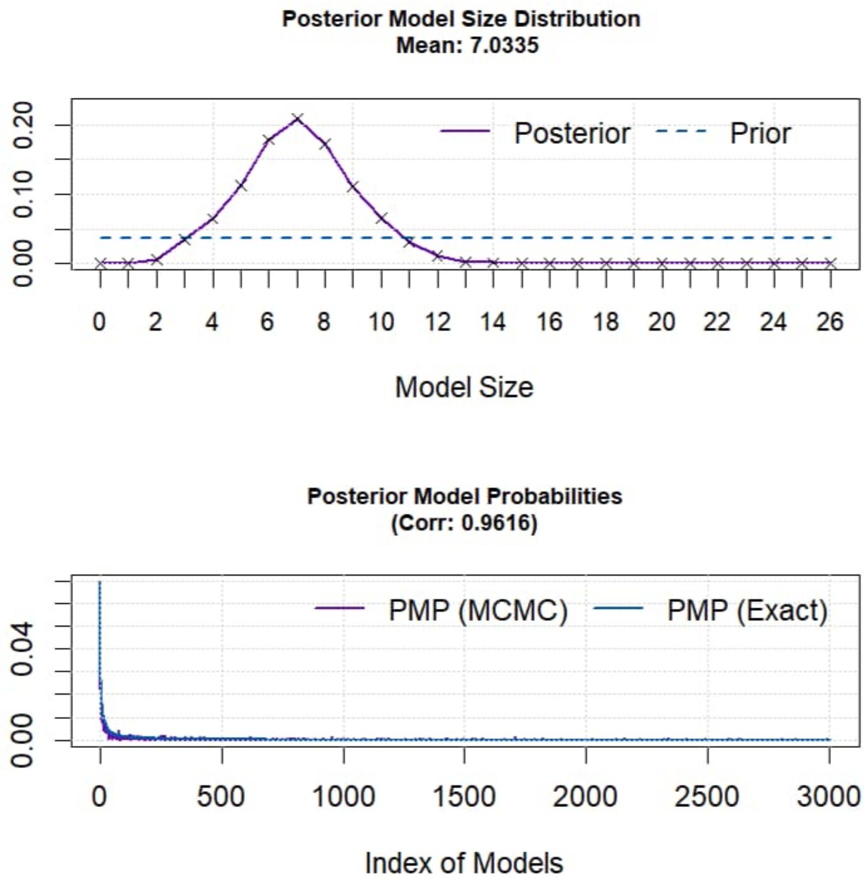
This section provides the reader with the results of Bayesian model averaging performed when excluding all studies performed by OECD, namely OECD (2001; 2003; 2004; 2005; 2007; 2010; 2014). We used the same setup as for the baseline BMA, see Chapter 5 for details.

Figure B.3: Correlation matrix of the explanatory variables for alternative BMA



Notes: This figure depicts the correlation table for our alternative BMA. Blue color denotes a positive correlation, while violet denotes a negative correlation. For description of the variables see Table 5.1.

Figure B.4: Posterior model size and convergence of the alternative BMA estimation



Notes: The figure presents the posterior model probabilities for different model sizes and posterior model size distribution of the alternative BMA.