CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Teachers Matter: A Meta-Analysis

Bachelor's thesis

Author: Marek Daňa

Study program: Ekonomie a Finance

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

Year of defense: 2024

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title. AI tools were used in this thesis to improve the writing style and grammar. The output generated by AI was used with respect to principles of academic integrity.

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

Prague, July 31, 2024

Marek Daňa

Abstract

Current literature suggests that more experienced teachers positively impact student results. The size of this effect varies across many studies. However, it has not yet been corrected for publication bias and model uncertainty. Through a comprehensive meta-analysis, this thesis explores the relationship between teacher experience and student achievement. I assemble a dataset of 131 estimates from 19 studies. Initial findings indicate an average 2% increase in test score standard deviation for each additional year of teacher experience. However, the presence of publication bias is evident, as demonstrated by linear tests and recently developed non-linear techniques. This thesis uses model averaging to investigate the influence of 21 variables on the teacher experience effect. After correcting for publication bias and applying the Bayesian model averaging method, the true effect of teacher experience in included studies appears nonexistent or indistinguishable from zero. Selective publication practices may have inflated positive effect reported previously.

JEL Classification	I21, H52, C83
Keywords	Teacher experience, test score, student achievement, education, meta-analysis, publication bias, Bayesian model averaging

Title Teachers Matter: A Meta-Analysis

Abstrakt

Současná literatura naznačuje, že zkušenější učitelé mají pozitivní vliv na výsledky studentů. Velikost tohoto účinku se mezi různými studiemi liší. Ovšem dosud nebyl započítán dopad publikačního zkreslení a nejistota modelu. Tato práce zkoumá vztah mezi zkušenostmi učitelů a úspěchy studentů prostřednictvím komplexní meta-analýzy. Sestavil jsem soubor dat obsahující 131 odhadů z 19 studií. Počáteční zjištění naznačují průměrné zvýšení standardní odchylky výsledků testů o 2% za každý další rok zkušeností učitele. Přítomnost publikačního zkreslení je však zřejmá, jak ukazují lineární testy a nově vyvinuté nelineární techniky. Pomocí průměrování modelů zkoumám vliv 21 proměnných na efekt zkušeností učitelů. Po započítání publikačního zkreslení a aplikaci metody Bayesovského průměrování modelů se zdá, že skutečný efekt zkušeností učitelů ve vybraných studiích neexistuje nebo není rozlišitelný od nuly. To naznačuje, že publikační zkreslení mohlo uměle zveličit pozitivní účinky zjištěné dříve.

Klasifikace JEL I21, H52, C83

Klíčová slova Zkušenosti učitele, skóre testu, výsledky studentů, vzdělání, metaanalýza, publikační zkreslení, Bayesovské průměrování modelů

Název práce Vliv učitelů: Meta-Analýza

Acknowledgments

I want to express gratitude, especially to my supervisor, Doc. PhDr. Zuzana Havránková, Ph.D. for the opportunity to write this thesis and her guidance. Further, I thank Mgr. Petr Čala for creating an open-source project with meta-analysis methods in R, which I used as inspiration and included several of his functions. I also want to thank prof. Tomáš Havránek for creating an insightful guide for meta-analysis.

Data and code are available here.

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
Daňa, Marek: *Teachers Matter: A Meta-Analysis*. Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 65. Advisor: Doc. PhDr. Zuzana Havránková, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

FAT-PET Funnel Asymmetry Tests - Precision Effect Testing

OLS Ordinary Least Squares

WAAP Weighted Average of the Adequately Powered

RoBMA Robust Bayesian Model Averaging

MAIVE Meta-Analysis Instrumental Variable Estimator

IV Instrumental Variable

BMA Bayesian Model Averaging

FMA Frequentist Model Averaging

HLM Hierarchical Linear Modeling

PIP Posterior Inclusion Probability

PMP Posterior Model Probability

Chapter 1

Introduction

The quality of education is crucial to societal advancement, shaping the capabilities of future generations and thus influencing economic, social, and cultural development. In the evolving landscape of educational research, the scrutiny of teacher quality and its direct correlation with student achievement occupies a central role in current research. Hanushek (2011) asserts that teachers represent the most influential factor in determining student achievement, with no other attribute of schools yielding a similar magnitude of impact. Among the myriad factors contributing to educational outcomes, the role of teacher experience has garnered considerable attention within academic circles.

As stated by De Paola (2009), an increase in teacher experience and teacher research productivity positively affects student performance. According to Graham *et al.* (2020), the mere accumulation of years of experience does not inherently result in superior quality of teaching. This discrepancy in current research is why I want to investigate the teacher-student relationship further. To my knowledge, no similar meta-analysis is available. The closest related topic is a meta-analysis suggesting that linking financial incentives to student outcomes has a positive and statistically significant effect (Pham *et al.* 2021). Another similar paper is Podolsky *et al.* (2019), which reviewed 30 studies, however, not using meta-analysis methods. Since studies report varying estimates, I will conduct a meta-analysis using the latest methods and techniques to evaluate reported estimates of the effect of teacher experience on student achievement. Even though student achievement might not seem as important, a paper by Rose (2006) showed that test score gains in high school forecast future employment and earnings. Even though the effect is limited for men, the effect is strong for women and men with test scores in the bottom quartile. It indicates

the economic significance and the possibility that this effect is not limited to test score gains only in high school.

I collected 131 estimates with their standard errors and test score standard deviations from 19 papers, following PRISMA standards. The details of this collection process are provided in Figure A.1. The most important criteria were that reported estimates needed to be accompanied by standard errors to calculate the precision of the effect and that test score standard deviation had to be available to normalize the comparison among studies. Firstly, I intended to include all indicators of teacher quality, such as education and training. However, data availability compelled me to focus only on causal estimates of the linear effect of teacher experience on student test score. A few studies focused on experience dummy variables, which did not assume a linear effect of experience.

In order to obtain a base knowledge of the effect size, I calculated both a simple mean and a mean weighted inversely by the amount of estimates included in each study. The simple mean indicated an increase in standard deviation of 0.02, while the weighted mean was 0.025. Therefore, before accounting for any biases in the reported estimates, one can assume that one additional year of teaching experience results in 2% increase in test score standard deviation. Assuming the linearity of this effect, having a teacher with an extra ten years of experience would result in a 20% increase in student test score standard deviation.

Publication bias, driven by the inclination of authors to favor findings that are both intuitive and statistically meaningful because publication is often based on these factors, poses a threat to the validity of reported estimates on teacher experience effects on students' test scores (Stanley 2005). Researchers often alter the specifications of models to achieve higher significance or manipulate the data to attain intuitive results (Gerber *et al.* 2008). There is a straightforward expectation based on intuition. More experienced teachers are presumed to enhance student learning or, at the minimum, not harm it. By applying the latest statistical techniques, I aim to correct the estimates for the effect of publication bias.

After observing the funnel plot to visually investigate the presence of publication bias (Stanley 2005), I implemented Funnel Asymmetry Tests - Precision Effect Testing (FAT-PET) with different specifications and weights such as Ordinary Least Squares (OLS), precision weighted by the inverse number of observations, or fixed effects. These tests rely on a simple regression equation,

where the effect size is the dependent variable and the standard error is the independent variable. Under strong assumptions (linear relationship, exogeneity) a significant estimate of the standard error coefficient indicates the presence of publication bias. The funnel plot suggested the possibility of publication bias, and five out of six linear tests confirmed the presence of publication bias. Effects beyond bias suggested by these tests range between -0.001 and 0.024, similar to mean results. However, it is evident that publication bias inflates the effect size.

Previous tests assumed a linear relationship between the effect and its standard error, which may not reflect reality. Therefore, I applied the following tests which do not rely on a linear structure of the relationship: The stem-based method (Furukawa 2019), Weighted Average of the Adequately Powered (WAAP) (Ioannidis *et al.* 2017), Top10 method (Stanley *et al.* 2010), the Endogenous kink model (Bom & Rachinger 2019), and the Selection model (Andrews & Kasy 2019). Top10 method should decrease the impact of publication bias, and it produced an insignificant effect. However, it might be related to a low number of estimates. The stem-based method also resulted in an insignificant estimate of the effect. The Endogenous kink model and Selection model clearly suggested publication bias.

To further test for publication bias, I turned my attention to the latest techniques, which do not operate under the exogeneity assumption. The first of these tests is FAT-PET with Instrumental Variable (IV) regression. The final choice of instrument was the inverse of the square root of the sample size. I continued with the p-uniform* method introduced by Van Aert & Van Assen (2021), the Meta-Analysis Instrumental Variable Estimator (MAIVE) recently developed by Irsova *et al.* (2023a), the Elliot tests by Elliott *et al.* (2022), and the Robust Bayesian Model Averaging (RoBMA) by Maier *et al.* (2022). The p-uniform* method does not suggest the presence of publication bias. However, Instrumental Variable Regression does indicate publication bias. Elliott tests do not indicate the presence of p-hacking in my dataset. Robust Bayesian Model Averaging presents strong evidence of publication bias. Unlike previous tests, there is no clear result when applying the latest techniques of publication bias detection.

Even though not every test indicated publication bias, generally, the findings support the presence of a moderate publication bias in my dataset. Effect beyond bias seems unclear, as different techniques produce varying results. There is a possibility that the effect is present only because of publication bias

since many methods produced an effect beyond bias indistinguishable from zero.

Since the effect size is influenced not only by publication bias, I have employed methods to address heterogeneity and and understand how various variables impact the size of the teacher experience effect. The use of divergent methods or different settings often causes discrepancies in studies. To address the issue of model uncertainty, I utilized both Bayesian Model Averaging (BMA) (Maier *et al.* 2022) and Frequentist Model Averaging (FMA) approaches (Steel 2020). These models allowed me to identify several important variables with a substantial impact on the effect size and many variables with negligible impact. For instance, the use of the fixed effects method has a very significant (positive) influence on the effect size, which is anticipated, as this method is used to mitigate endogeneity (Clotfelter *et al.* 2010) and the fact that teachers are not randomly distributed across classrooms. Adding teacher fixed effects into the model is the most common solution to reduce bias and obtain most accurate estimates of how much teachers improve while they gain more years of teaching experience (Podolsky *et al.* 2019).

Results also indicate that the impact of teacher experience differs based on the grade levels of the students. In the Appendix, I also included a robustness check of Bayesian model averaging using different specifications, specifically a uniform g-prior with a uniform model prior and an HQ g-prior with a random model prior. However, the results from these different specifications are all very similar, confirming the robustness of the findings.

This meta-analysis does not produce a positive estimate of the effect size of one additional year of teacher experience on student test scores. Many results are not significantly different from zero, and many provide conflicting results ranging from slightly negative to considerably positive. Together, the papers included in this meta-analysis do not provide convincing evidence of the existence of teacher experience effect on student test score beyond publication bias. The effect does not exist or is not distinguishable from zero. Current research must focus on discovering other meaningful variables that indicate teacher quality.

The structure of this thesis is as follows: Chapter 2 investigates available literature on the teacher experience effect. Chapter 3 describes data collection and the construction of the dataset alongside basic numerical and graphical analysis of collected estimates. Chapter 4 examines publication bias using modern methods. Chapter 5 applies model averaging techniques to observe

differences between studies and explain heterogeneity. Chapter 6 presents the best-practice estimates based on the author's subjective view and the entire dataset. Chapter 7 summarizes the thesis.

Chapter 2

Teacher experience

A fundamental approach to studying teacher effects begins with an achievement model, where the outcome for a student in grade g is a cumulative function of vectors representing family, teacher, school, and ability. Simply put, student achievement is a result of many different inputs. This model, commonly known as an educational production function, is widely used (Hanushek *et al.* 2004). However, it faces the risk of bias from many sources. For example, as shown by Hanushek *et al.* (2004), teachers tend to move to schools with higher student achievement, creating the possibility of a simultaneous equation bias. The expected direction of causation is that teacher experience increases student achievement; however, Hanushek *et al.* (2004) suggests student achievement increases teacher experience (Hanushek & Rivkin 2006). The complexity of education makes it difficult to obtain unbiased and precise estimates.

One can distinguish two effects of teaching experience. The first one, which this thesis focuses on, is the return to teacher experience, which means how much teacher improve over the years by gaining experience. The second effect is the chance that teachers with more experience are simply those with better skills who were able to stay in the teaching profession. Similarly, teachers are not randomly assigned to students within the school; teachers with more experience usually teach students with better abilities, which causes upward bias. (Podolsky *et al.* 2019).

Since I normalize test scores using standard deviation, it is essential to provide a tangible reference to make the results more accessible. Cremata *et al.* (2013) provides a general approximation, where 0.02 growth in standard deviations corresponds to an additional 14 days of learning.

Chapter 3

Data

3.1 Literature

To investigate the influence of teacher quality on student achievement, I leveraged the comprehensive full-text search capabilities of Google Scholar. The primary strength of Google Scholar is the power to search the whole content of papers, rather than being limited to titles, abstracts, and keywords. This feature is particularly beneficial for ensuring that relevant studies are not overlooked. The search strategy, detailed in the PRISMA flow diagram along with the particular search query, is reported in Figure A.1 in the Appendix. The query was precisely crafted to capture the most relevant and impactful studies on the effect of teacher quality on student achievement. The focus was strictly on studies published in the English language. The search was conducted in April 2024, followed by a snowballing process in May 2024 to expand the dataset further.

The snowballing process began with 17 studies identified through the initial Google Scholar query, which had already undergone a rigorous selection process. To expand the dataset, I used Scopus to gather references from these studies and scrutinized the most frequently cited ones. This method yielded 17 promising studies, of which only two ultimately passed the selection criteria.

Initially, I screened abstracts of the first 500 papers returned by Google Scholar search, retaining 214 papers that appeared relevant. I repeated the search to ensure up-to-date coverage, focusing on the first 30 papers published in the last three years. After reviewing the abstracts of these 530 papers, 210 were deemed to contain potentially valuable data for further analysis. The snowballing method contributed 17 papers that could have provided estimates

for inclusion in the dataset. In total, 227 studies were assessed in detail, and 19 were included in the meta-analysis, providing 131 estimates of the effect. According to Irsova *et al.* (2023b), a robust meta-regression analysis requires at least 30 estimates of the effect derived from a minimum of 10 studies. The complete list of studies included in the meta-analysis is provided in Table 3.1.

Each study had to meet three specific criteria for inclusion in the meta-analysis. First, it must report one or more estimates of the relationship between test scores and teacher experience, along with standard errors, t-statistics, or other information that can be transformed into standard errors. Provision of standard errors is essential to assess the precision of the effect across different observations. Second, studies needed to report the standard deviations of test scores. This information is crucial for standardizing the effect, allowing it to be converted to a common metric. Without standardization, the effect size would lack comparative value, as different studies use different tests, resulting in varying score scales. The use of standard deviations addresses this issue by normalizing the scores. In the dataset, an estimate of 0.01 indicates that a one-year increase in teacher experience is associated with an improvement in test scores of 0.01 standard deviation. Third, each estimate had to be accompanied by an identifiable statistic indicating the sample size. If these statistics were not provided, the study had to specify the number of students involved—the sample size needed to correspond accurately to the reported estimates to ensure the validity of the data.

To maximize the number of estimates, I decided not to exclude any estimates based on the quality of the studies. Overall, there is a rationale for including all studies that meet the inclusion criteria. This approach enables the identification of how variations in methodology affect the results, which may be the primary reason for conducting the meta-analysis (Irsova *et al.* 2023b).

3.2 Construction of dataset

Apart from collecting 131 estimates from 19 relevant papers on the effect of one year of teacher experience on test score standard deviation, I gathered other factors to encapsulate heterogeneity within the literature. I collected 26 variables, including variables capturing data characteristics, estimation technique, and publication characteristics. A complete list of collected variables, including their description and mean value, can be found in Table 5.1. Regarding publication characteristics, my dataset does not reflect any changes after May 2024.

Table 3.1: Studies obtained for purpose of meta-analysis

Goldhaber & Brewer (1996)	Betts & Shkolnik (2000)
Goldhaber & Brewer (2000)	Hill *et al.* (2005)
Borman & Kimball (2005)	Jepsen (2005)
Darling-Hammond *et al.* (2005)	Krieg (2006)
Kukla-Acevedo (2008)	Munoz & Chang (2007)
Miller *et al.* (2008)	Sutton & Soderstrom (1999)
Kingdon & Teal (2010)	Leigh (2010)
Reeves *et al.* (2016)	Blazar (2015)
Canales & Maldonado (2018)	Penner (2021)
Sancassani (2023)	

To reduce human error in the manual coding of studies, two authors should collect the data independently (Irsova *et al.* 2023b). Since this was impossible for this thesis, I did the coding twice and verified no differences in results. To mitigate any mistakes the authors of papers might have created, I winsorized the effect, standard error, and t-stat at the 2.5% level from each side.

3.3 Analysis of dataset

Table 3.2 and Table 3.3 show the mean effect of teacher experience conditionally on certain variables. The first table reports unweighted statistics, where each estimate is given equal weight (simple mean). The second table displays the weighted mean, where the weight is inverse to the amount of estimates that are included in each study, ensuring that each study contributes equally to the overall results. We can observe significant differences between subsets. When we consider all data, the mean effect is positive, indicating around 0.02 standard deviation increase in test scores for one additional year of teacher experience. The weighted mean effect is even more significant at 0.025 standard deviation.

The methodology utilized significantly influences the mean effect, with estimates being substantially higher when employing the fixed effects method than Ordinary Least Squares (OLS). This effect is notably more prominent than that observed with the Random Effects and Hierarchical Linear Modeling (HLM) methods, which demonstrate a negative effect; however, only 12 observations are available for these two methods. Although the difference is not as pronounced for the weighted mean, it remains noticeable.

The mean effect is more pronounced in mathematics tests compared to reading or other subjects, suggesting that the impact of teacher experience

may vary by subject. The student sample also appears to affect the mean effect size, with students in the lower grades, particularly from kindergarten to 5th grade, being more significantly influenced by teacher experience. Most of the analyzed studies were conducted in the United States, where the mean effect is more substantial. This discrepancy could be attributed to differences in the educational system or the precision of the research conducted.

Although these results suggest a positive trend in the effect, they should not be considered definitive, and one should approach all these findings with skepticism. Firstly, the precision is limited due to quite wide size of the confidence intervals. Secondly, even if weighted, the mean is not a perfect measure, and other statistical methods should be employed. Further analysis of heterogeneity is conducted in Chapter 5.

To underline the distinctions between individual studies, a box plot of estimates at the study level is included in Figure 3.1 and Figure 3.2. The studies are arranged chronologically, from the oldest to the most recent. Due to the differences in the size of reported estimates, I have divided the data into two plots. Specifically, two studies, Miller *et al.* (2008) and Kukla-Acevedo (2008), reported estimates up to 0.5, whereas the other seventeen reported estimates up to 0.027. This discrepancy would affect the ratio and visibility of the smaller box plot. Suspecting a coding error, I thoroughly reviewed the data collection process. The review showed no errors, so I confirmed that the data were error-free. For clarity, extreme outliers were removed but were part of all statistical tests.

The box plot reveals that while some studies indicate a negative effect, most show an effect slightly above zero, with a few studies even suggesting a substantial positive impact of teacher experience. Crucial to remember is the fact that the effect has been normalized to the test score standard deviation, making the effect appear smaller. Additionally, even within individual studies, the reported estimates vary significantly.

Table 3.2: Mean effect of teacher experience for certain factors — Unweighted

	Mean	95% CI	Observations
All Data	0.0198	(0.0078, 0.0318)	131
Sample characteristic			
Study conducted in USA	0.0278	(0.0110, 0.0446)	92
Study conducted in other country	0.0010	(-0.0001, 0.0021)	39
Sample from kindergarten to 5th grade	0.0372	(0.0145, 0.0599)	67
Sample from 6th to 8th grade	0.0005	(0.0000, 0.0111)	29
Sample from 9th to 12th grade	0.0033	(0.0004, 0.0062)	28
Sample from different grade	-0.0011	(-0.0013, 0.0008)	7
Test characteristics			
Math test	0.0401	(0.0157, 0.0645)	62
Reading test	0.0007	(-0.0003, 0.0016)	26
Test in other subject	0.0021	(0.0002, 0.0040)	42
Method characteristics			
FE method used	0.0401	(0.0138, 0.0665)	57
OLS method used	0.0047	(-0.0006, 0.0101)	50
RE method used	0.0062	(0.0009, 0.0116)	12
HLM method used	-0.0003	(-0.0011, 0.0005)	12

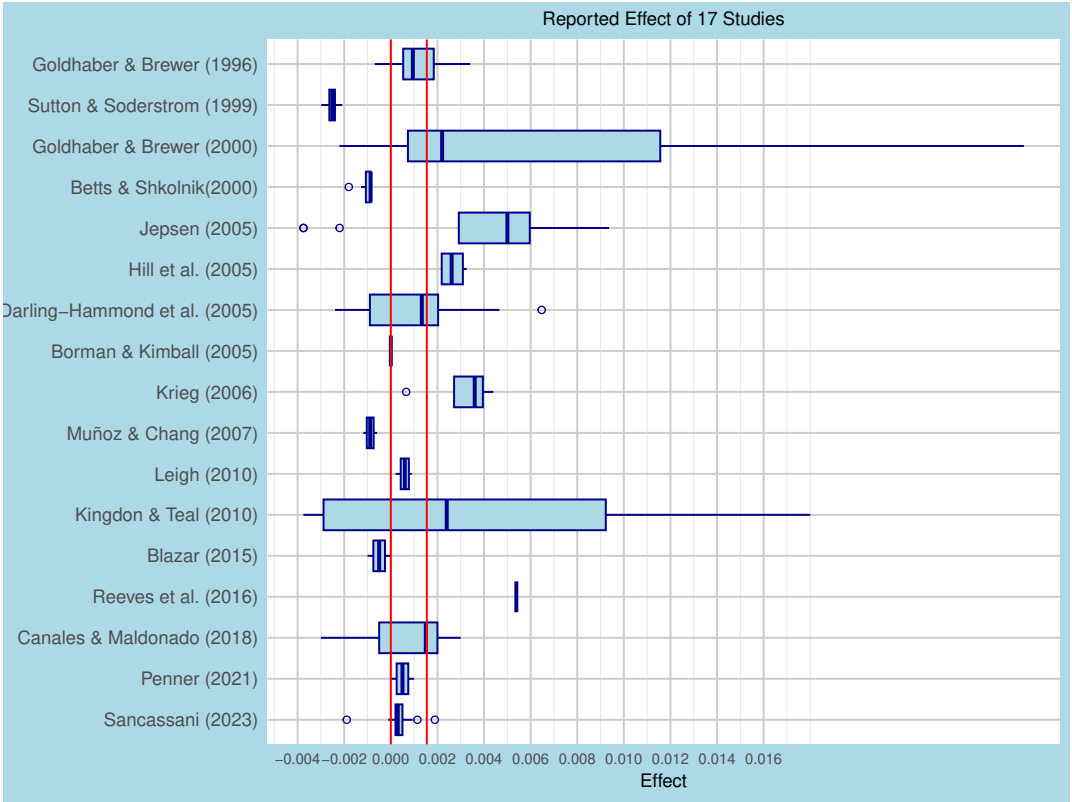
Note: The table reports summary statistics for the effect of one year of teacher experience on test scores (normalized by standard deviation) calculated conditionally for different variables. 'Unweighted' indicates that data are used without modifications, therefore just simple mean. CI = Confidence interval, FE = Fixed effects, OLS = Ordinary Least Squares, RE = Random effects, HLM = Hierarchical linear modeling.

Table 3.3: Mean effect of teacher experience for certain factors — Weighted

	Mean	95% CI	Observations
All Data	0.0256	(0.0219, 0.0293)	131
Sample characteristic			
Study conducted in USA	0.0323	(0.0269, 0.0376)	92
Study conducted in other country	0.0007	(-0.0005, 0.0020)	39
Sample from kindergarten to 5th grade	0.0557	(0.0441, 0.0673)	67
Sample from 6th to 8th grade	0.0016	(0.0005, 0.0027)	29
Sample from 9th to 12th grade	0.0016	(0.0005, 0.0027)	28
Sample from different grade	-0.0011	(-0.0013, 0.0008)	7
Test characteristics			
Math test	0.0447	(0.0358, 0.0535)	62
Reading test	-0.0004	(-0.0011, 0.0003)	26
Test in other subject	0.0019	(0.0007, 0.0031)	42
Method characteristics			
FE method used	0.0497	(0.0376, 0.0619)	57
OLS method used	0.0104	(0.0051, 0.0156)	50
RE method used	0.0054	(0.0024, 0.0084)	12
HLM method used	-0.0004	(-0.0008, 0.0000)	12

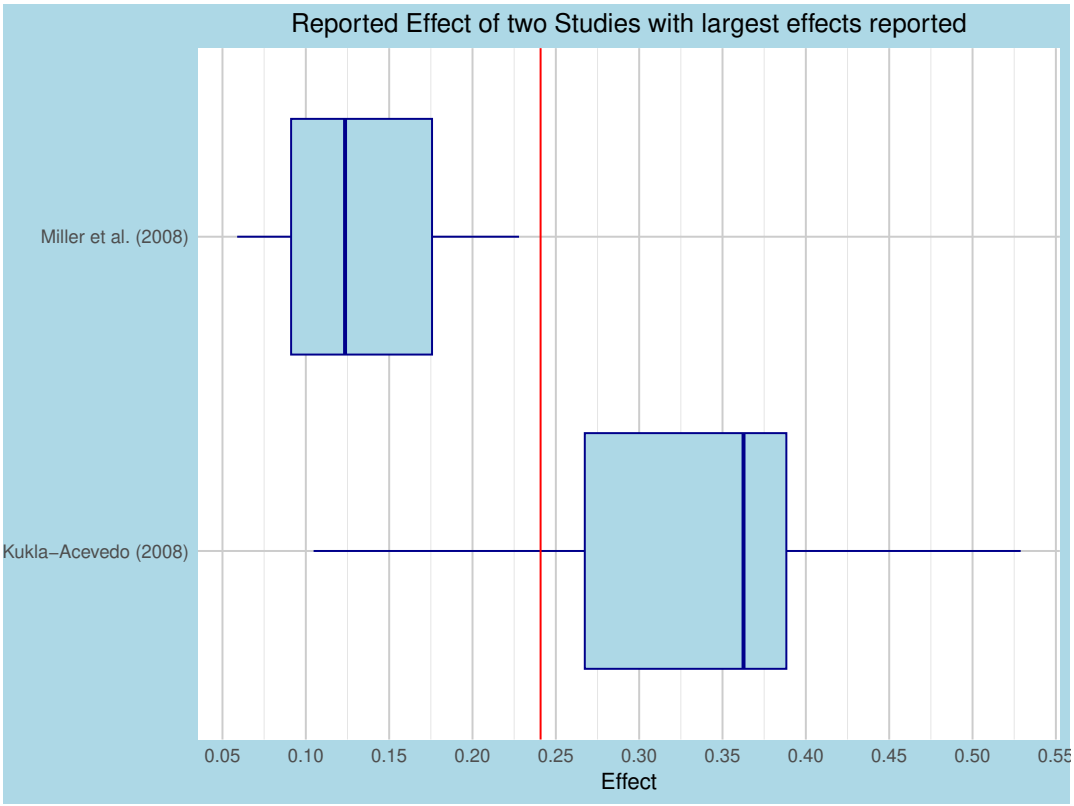
Note: The table reports summary statistics for the effect of one year of teacher experience on test scores (normalized by standard deviation) calculated conditionally for different variables. 'Weighted' indicates that the estimates are adjusted by weighting them inversely to the amount of estimates that are included in each study. CI = Confidence interval, FE = Fixed effects, OLS = Ordinary Least Squares, RE = Random effects, HLM = Hierarchical linear modeling.

Figure 3.1: Box plot of reported estimates in 17 studies



Note: The figure presents a box plot representing teacher experience effect estimates in 17 papers. Studies are ordered by year from oldest to newest. Estimates represent one standard deviation increase in student test scores based on a one-year increase in teacher experience. Length of each box depicts the distance between the first quartile and the third quartile, while the line in the box shows the median value. The whiskers spread to the most extreme points of data within 1.5 times the interquartile range from the lower and upper quartiles. The first red line indicates size zero. The second red line represents the mean effect size. For clarity, extreme outliers were removed but incorporated in every statistical test.

Figure 3.2: Box plot of reported estimates in 2 studies



Note: The figure presents a box plot representing teacher experience effect estimates in 2 papers. Studies are ordered by year from oldest to newest. Estimates represent one standard deviation increase in student test scores based on a one-year increase in teacher experience. Length of each box depicts the distance between the first quartile and the third quartile, while the line in the box shows the median value. The whiskers spread to the most extreme points of data within 1.5 times the interquartile range from the lower and upper quartiles. The first red line indicates size zero. The second red line represents the mean effect size. For clarity, extreme outliers were removed but incorporated in every statistical test.

Chapter 4

Publication Bias

A significant challenge in understanding research about teacher experience and student achievement is the issue of publication bias or p-hacking. This bias occurs when studies with specific results have higher chance of being published. According to Stanley (2005), publication selection bias typically arises when only findings that are statistically significant or conform to prevailing theories are published. This bias is usually associated with a correlation between estimates and their standard errors.

The definitions of publication bias and p-hacking can vary significantly. Publication bias is sometimes broadly defined as all instances where the reported research findings differ from the initial results obtained by the authors when they first analyze their data. In another definition, publication bias is related to the problem of unpublished studies with non-significant or unexpected results. At the same time, p-hacking is the deliberate or subconscious manipulation with data or analytical techniques to achieve statistical significance. In actual data, these two occurrences can appear identical to a meta-analyst; therefore, I will use the terms interchangeably. (Irsova *et al.* 2023b)

As per Podolsky *et al.* (2019), teaching experience is positively linked with student achievement results throughout a teacher's career. The most significant gains occur in the initial years of teaching, but significant improvements continue into the second and frequently third decades of a career. There is a consensus that teacher experience positively affects test scores, which may create a publication bias when researchers obtain opposite results. Publishers might refuse to publish findings that contradict common knowledge or are statistically insignificant.

Having invested significant time in their studies, researchers might modify

the specifications to obtain more acceptable or significant results. This discrimination against counterintuitive results leads to what is known as the file drawer effect (Stanley 2005). Modifying results to achieve higher significance of estimates is often called p-hacking. (Irsova *et al.* 2023b)

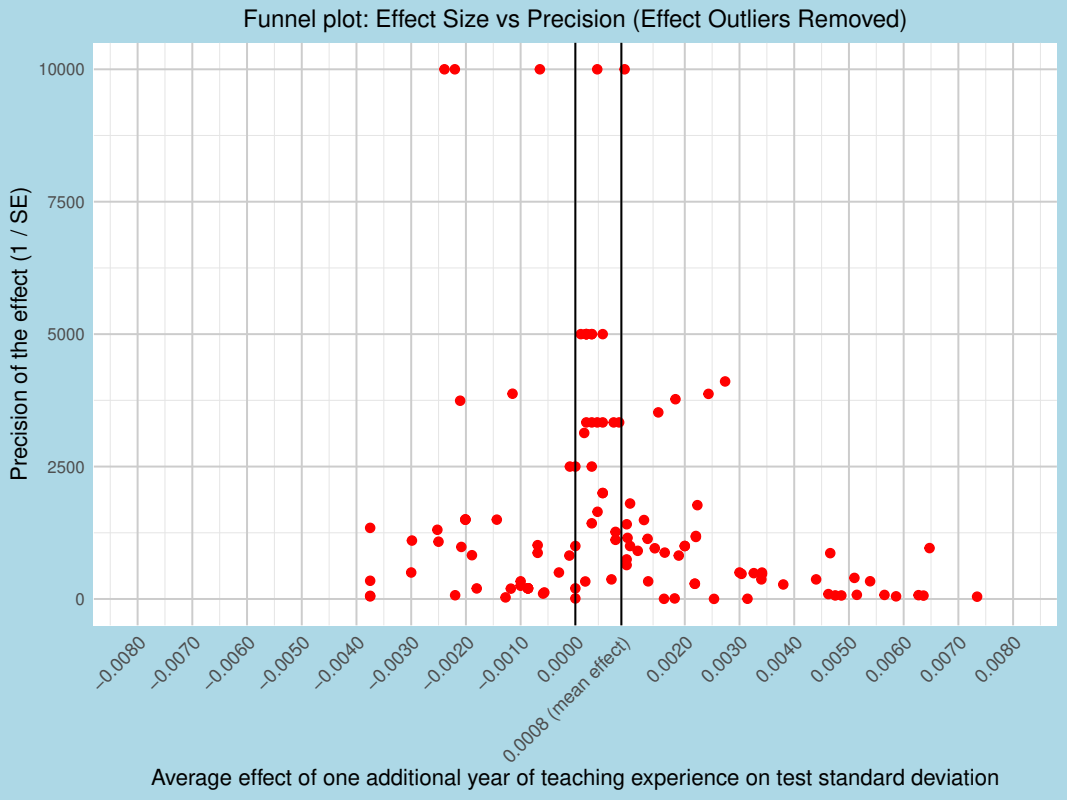
In the remainder of this segment, I will utilize a straightforward graphical test to visually investigate whether the effect is associated with the standard error. Many statistical methods can detect publication bias within data. I will execute linear and non-linear tests to decide if publication bias is present in my dataset. Next, I will conduct tests that do not require the exogeneity assumption. Altogether, I can determine whether my dataset is influenced by p-hacking.

4.1 Funnel Plot

Firstly, I will utilize a visual method, a funnel plot Figure 4.1. Developed by (Egger *et al.* 1997), this plot is a scatter diagram where the size of the study effect is plotted horizontally against its precision (inversely related to the standard error) vertically. Stanley (2005) suggests other precision variants, for example, using some form of degrees of freedom. However, since my dataset has standard errors, I will use the most common version of the funnel plot.

In an ideal scenario, the most accurate studies would cluster at the top center of the chart, representing the true average effect, with less precise studies dispersing towards the bottom. Without publication bias, a symmetrical inverted funnel shape would be created. Any asymmetry in the funnel plot might represent an inclination in published papers to report overly positive or negative results, thus suggesting publication bias. However, Egger *et al.* (1997) mentions other sources of asymmetry such as heterogeneity or data anomalies.

Figure 4.1: Funnel plot



Note: The figure shows the funnel plot developed by Egger *et al.* (1997), the increase in test score standard deviation based on one year increase in teacher experience is plotted on the x-axis, while precision is plotted on the y-axis. A symmetrical funnel plot centered around the true mean would be an indicator of no publication bias present. The first black line shows size zero. The mean effect is depicted as the second black line.

The plot may have gaps due to the lower observation count of 131 estimate points, and asymmetry could be misleading. As expected, estimates are widely spread at the bottom and cluster around the mean. The funnel plot appears slightly asymmetrical on the positive side, which could be anticipated due to the strong intuition regarding the positive effect of teacher experience on test scores. At the highest levels of precision, a few data points are unexpectedly dispersed. Overall, there is some evidence of p-hacking, but additional tests are needed to confirm this. Even though funnel plot is a simple and quick method, it relies on subjective evaluation using only one's eyes.

4.2 Linear tests

To formally establish whether publication bias is present in the data, I focus on Funnel Asymmetry Tests - Precision Effect Testing (FAT-PET) (Egger *et al.* 1997; Stanley & Doucouliagos 2015). These tests use a simple regression equation with the effect as the outcome variable and standard error as the explanatory variable to discover any relationship between them. If there exists any significant connection between the two, it indicates the presence of publication bias.

$$T_{ij} = \beta_0 + \beta_1 \cdot (SE_T)_{ij} + \epsilon_{ij}, \tag{4.1}$$

In this Equation 4.1, T denotes the teacher experience effect, and (SE_T) indicates the standard error associated with this effect. Index i represents the observation number, and index j corresponds to the study number in the dataset.

The constant coefficient, β_0, represents the effect of teacher experience on the test scores after adjusting for publication bias. The slope coefficient, β_1, quantifying the relationship between the effect size and its standard error, measures the extent of publication bias in the dataset. The term ϵ_{ij} should be regarded as the error term in the regression analysis. The term 'Effect beyond bias' denotes the effect adjusted for publication bias or constant coefficient (β_0). In contrast, 'Publication bias' represents the slope of the relationship or the size of publication bias (β_1).

If the dataset exhibits no publication bias, the slope coefficient β_1 will be zero or near zero. Conversely, higher absolute values of the β_1 coefficient would indicate the presence of publication bias, evidenced by a significant correlation between the effect size and its standard error. This analysis assumes that the effect size and standard error are drawn from an independent, statistically symmetrical distribution. This assumption is often violated in practice (Andrews & Kasy 2019).

The results can be observed in Table 4.1. The first model is a simple Ordinary Least Squares (OLS) regression, followed by two models, "STUDY" and "PRECISION", which introduce weights into the regression; weight is calculated either as the inverse amount of estimates in each study or as precision (1/SE, Standard Error), respectively. Fixed Effect and Random Effect models are employed to address unobserved heterogeneity. Lastly, the Between Effects model is included, obtained by regressing the averaged variables on each other using

OLS regression. This model can be viewed as an intermediate step between the Fixed Effects and Random Effects models, giving each study the same weight. Standard errors are clustered at the study level, and wild bootstrap confidence intervals with 100 iterations were used where feasible.

Based on the results, five out of six models found statistically significant evidence of publication bias. However, there are substantial differences in the size of the β_1 coefficient across models, ranging from 0.2807 to 1.0996. Only the Fixed Effects model did not find the presence of publication bias. The effect cleared of publication bias is significant in three models. The Fixed Effects and Random Effects models show effects of 0.0179 and 0.0101, respectively, slightly lower than the simple average from Table 3.2. However, the 'Precision' model shows a small negative effect. Even though Publication Bias and Effect Beyond Bias are significant in the 'Precision' and Random Effects models, the values differ substantially. Since only some results are statistically significant and show varying values, further tests are required to determine which models are closest to reality.

Table 4.1: Publication bias detection — Linear tests

	OLS	**STUDY**	**PRECISION**
Publication Bias	0.3838***	0.4190***	1.0996**
(Standard error)	(0.0190)	(0.0204)	(0.3934)
Bootstrap CI	[0.2547, 0.4779]	[0.3033, 0.5235]	[0.2316, 1.8105]
Effect Beyond Bias	0.0004	0.0025	-0.0005***
(Standard error)	(0.0032)	(0.0033)	(0.0001)
Bootstrap CI	[-0.0021, 0.0032]	[-0.0039, 0.0100]	[-0.0012, 0.0006]
Observations	131	131	131
	Fixed Effects	**Between Effects**	**Random Effects**
Publication Bias	0.0387	0.4267***	0.2807***
(Standard error)	(0.0384)	(0.0421)	(0.0294)
Bootstrap CI			[0.0632, 0.4314]
Effect Beyond Bias	0.0179***	0.0021	0.0101*
(Standard error)	(0.0028)	(0.0066)	(0.0061)
Bootstrap CI			[-0.0006, 0.0238]
Observations	131	131	131

Note: The table reports the results of following estimation: $T_{ij} = \beta_0 + \beta_1 \cdot (SE_T)_{ij} + \epsilon_{ij}$, where T_{ij} denotes the i-th teacher experience effect estimated in the j-th study, and $(SE_T)_{ij}$ denotes the standard error. CI = Confidence Interval, OLS = Ordinary Least Squares. 'PRECISION' and 'STUDY' refer to estimates weighted inversely by the standard error and by the inverse number of estimates reported per study, respectively. Standard errors are clustered at the study level. Wild bootstrap 95% confidence intervals, created over 100 iterations. ***p<0.01, **p<0.05, *p<0.1

4.3 Nonlinear tests

In previous analyses, we assumed a linear relationship between the effect and its standard error, which can be problematic as it does not necessarily reflect real-world practice. Due to the weak or inconclusive results of these previous linear tests, I will employ the following non-linear tests: The stem-based method (Furukawa 2019), WAAP (Ioannidis *et al.* 2017), Top10 method (Stanley *et al.* 2010), Endogenous kink model (Bom & Rachinger 2019), and Selection model (Andrews & Kasy 2019). These techniques do not require the potentially unrealistic assumption of linearity between effect size and standard error. However, they do assume that estimates and standard errors are uncorrelated in the absence of bias.

The Weighted Average of the Adequately Powered model employs a funnel plot to exclude estimates with power under 80%. The power is obtained by contrasting significance to standard error. Then, the model calculates an mean weighted by inverse variance of estimates that fulfill the 80% rule.

The Top10 method involves removing 90% of the data, retaining only the estimates with the highest precision, specifically those in the top 10 percent. According to Stanley *et al.* (2010), this approach significantly reduces publication bias in the data sample.

The Stem-based method builds on these approaches by endogenously determining the share of the most significant estimates to utilize. It selects an amount of the most informative estimates by balancing the tradeoff between bias and variance: discarding estimates increases variance (inefficiency) while including imprecise estimates increases the risk of selective reporting (publication bias). The stem-based technique aims to limit the combined effect of bias and variance.

The endogenous kink technique modifies the Egger regression by incorporating a stable part of estimates that are the most statistically significant based on the assumption that publication bias is unaffected by changes in standard error for these estimates. Bom & Rachinger (2019) employed a piecewise linear regression with a kink at the cutoff point, introducing non-linearity into the model (Cala 2024). The main benefit of method developed by Bom & Rachinger (2019) is that it acts as linear when the effect is close to zero, which is an area where linear methods perform the best.

In the selection model, the authors (Andrews & Kasy 2019) propose that the probability of publication is constant at comparable degrees of statisti-

cal significance of estimates, a notion referred to as 'conditional publication probability'. When an estimate is above a specific significance threshold, its publication probability differs from publication probability of estimate that is under the threshold. This model estimates the probability of publication for each estimate within specific significance brackets and weights the reported estimates inversely to these probabilities. The authors provide method of non-parametric calculation of publication probability. Authors achieve an unbiased distribution of the estimates by using new weights equal to inverse of this probability. Described approach utilizes a t-distribution with a 5% significance level, setting the cutoffs at 1.96. (Cala 2024)

Table 4.2 presents the numerical results of the previously explained methods. Unfortunately, for the WAAP method, none of the estimates met the 80% power level. After mitigating publication bias, the Top10 method produced an insignificant estimate of the teacher experience effect. The stem-based method also resulted in insignificant findings. However, using a selection model with a t-distribution of the 5% significance level, significant estimates were obtained, indicating publication bias and a true effect of teacher experience of 0.00058, considerably smaller than the results of previously conducted linear tests. The endogenous kink model also suggests the presence of publication bias, as indicated by a significant estimate, but it suggests a negative true effect of -0.00045. It is important to note that non-linear techniques corrected the effect beyond bias closer to zero, which points to the presence of publication bias in my dataset.

Table 4.2: Publication bias detection — Nonlinear tests

	Stem	WAAP	Top10%	Kink	Selection
Publication Bias				1.09960**	0.00129***
(Standard error)				(0.39343)	(0.00020)
Effect Beyond Bias	0.00038		-0.00015	-0.00045***	0.00058*
(Standard error)	(0.00046)		(0.00027)	(0.00015)	(0.00025)
Model observations	115	0	13	131	131

Note: This table presents the results of five non-linear methods to detect publication bias. Stem = the Stem-based method Furukawa (2019). WAAP = Weighted Average of the Adequately Powered Ioannidis *et al.* (2017). Top10% = Top10 method Stanley *et al.* (2010). Kink = Endogenous kink model Bom & Rachinger (2019). Selection = Selection model by Andrews & Kasy (2019). The publication bias column corresponds to the detection of publication bias and Effect Beyond Bias reports the size of the effect of teacher experience after incorporating publication bias. Standard errors are clustered at the study level. ***p<0.01, **p<0.05, *p<0.1

4.4 Endogeneity-robust techniques

In previous sections, we used techniques built that assume that any relationship linking effect and its standard error is evidence of publication bias. Under that assumption, the endogeneity issue is automatically created in the equation as the standard error is influenced by the same factors driving the effect size. Endogeneity and correlation can have various roots. Endogeneity might stem from mistakes in measurement or inaccurate methods of calculation, considering that the standard error is also an estimate. Further, endogeneity is often associated with publication bias, which may arise from unintentional or deliberate manipulation of the standard error to get statistically significant results. Lastly, unobserved heterogeneity can introduce correlation due to choices in study design. For example, choosing a particular regression model or using a specific sample could affect the effect size. This, of course, applies to standard errors as well. (Cala 2024)

First, I utilize IV regression. For this, a suitable instrument must be found. It should be a variable correlated to standard error while not being expected to have any connection to the effect. This instrument should be capable of providing publication bias coefficient β_1 not affected by endogeneity.

Many forms of such an instrument could be utilized, including $\frac{1}{n_{\text{sample}}}$, $\frac{1}{\sqrt{n_{\text{sample}}}}$, n_{sample}, $\frac{1}{n^2_{\text{sample}}}$, and $\log(n_{\text{sample}})$, where n_{sample} represents the sample size from

which reported estimates were obtained. Sample size works the best as it does not directly change the teacher experience effect since it is independent of sample size. Another aspect is that the standard error decreases with a bigger sample size; therefore, the effect is more precise.

I will utilize the inverse of the square root of the sample size. It has provided the most valuable insights. It is essential to recognize that many tests require a substantial number of observations to yield reliable results. (Havranek *et al.* 2022). Instrumental Variable Regression suggests the presence of publication bias since it produced a significant estimate of 0.4805. However, effect size beyond publication bias is undetermined as its estimate is insignificant.

The second method used is p-uniform* (Van Aert & Van Assen 2021); it is based on the statistical theory that the distribution of p-values is uniform and conditional on the true effect size. This approach does not require assumptions about the form or correlation of the relationship, as it tests publication bias utilizing the distribution of every p-value, offering a unique way to detect publication bias. The p-uniform* method does not suggest the presence of publication bias with a p-value of 0.721 and does not give a significant estimate of effect size. It is a surprising result, as many previous tests indicated the existence of publication bias in this dataset.

Table 4.3: Endogeneity-robust techniques

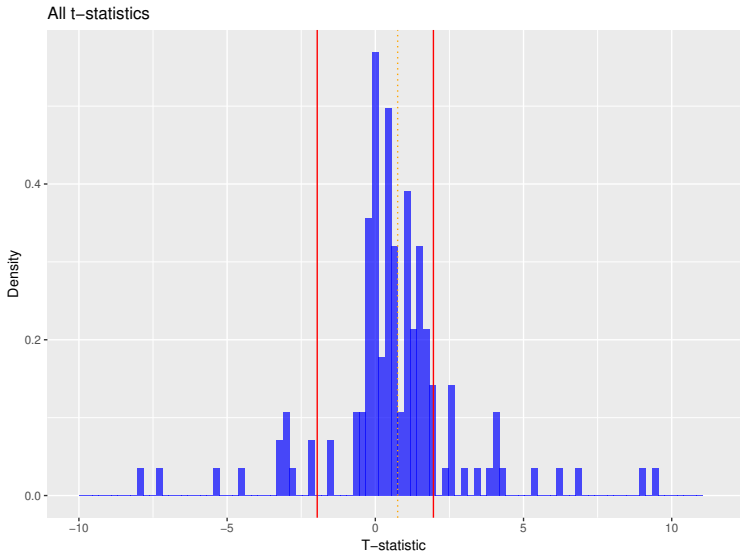
	IV	**p-uniform***
Publication Bias	0.4805**	L = 0.1279
(Standard error)	(0.1990)	(p = 0.721)
Effect Beyond Bias	-0.0045	0.0050
(Standard error)	(0.0095)	(0.0135)
Observations	131	131
AR	0.7305	

Note: IV = IV regression; as an instrument is used invers of the square root of the sample size. p-uniform* = method proposed by Van Aert & Van Assen (2021); L describes the publication bias test t-statistic with p-value in parentheses. AR = Anderson-Rubin F-test statistic Anderson & Rubin (1949). Standard errors are clustered at the study level. ***p<0.01, **p<0.05, *p<0.1

4.5 Caliper test

Another approach to avoid exogeneity assumption was developed by Gerber *et al.* (2008), namely the Caliper test. Unlike previous techniques, it is focused on the distribution of t-statistics instead of the relationship between effect size and standard error. The core idea is that publication bias explains unexpected jumps in the t-statistic distribution. For example, one can take t-statistic 1.645 and examine whether the number of t-statistics just above this level, within some small interval, is approximately the same as t-statistics just under this level. Any significant difference in the number of t-statistics just above against just under is an indicator of structural break, suggesting the presence of publication bias. On the other hand, in the absence of bias, the t-statistic distribution should be normally distributed without sudden jumps. However, based on Figure 4.2, I decided not to include the test in this thesis because a small number of observations made the results inconclusive.

Figure 4.2: Histogram of t-statistic



Note: The figure plotted the distribution of estimates t-statistics. Red lines denote the significant t-values -1.96 and 1.96. The distribution of t-statistic is skewed as expected. Outliers excluded for clarity of the graph were included in the computations.

4.6 Further Detection of Publication Bias

In this chapter, I will utilize the latest techniques to detect the presence of publication bias. These methods, developed in recent years, represent the state of the art. Specifically, this chapter includes the following: The Meta-Analysis Instrumental Variable Estimator (MAIVE) by Irsova *et al.* (2023a) and Elliot tests by Elliott *et al.* (2022).

MAIVE emphasizes spurious regression and how p-hacking causes endogeneity in results. Irsova *et al.* (2023a) offers an elegant solution to measure the reported standard error. MAIVE employs the inverse square root of the sample size as an instrument. Even after applying this solution, some endogeneity is still left, but endogeneity is reduced significantly. It is generally much more challenging to alter sample size compared to shrinking standard error and it does not depend on measurement error or estimation methods (Opatrny *et al.* 2023). Table 4.4 offers results of MAIVE, the effect of teacher experience is statistically insignificant, and low F-statistic 0.340 suggests inverse sample size is not a strong instrument for standard error reported.

Elliott tests inspect the distribution of p-values among studies and propose an approach that considers the absence of p-hacking in the dataset as the null hypothesis. Under a set of broad assumptions, they test absence of p-hacking hypothesis against the alternative hypothesis that p-hacking is present in the dataset. Elliot tests use two techniques. Both techniques are histogram-based, one tests for the non-increasingness of the p-curve and the other for monotonicity and bounds on the p-curve. The original specification used by Elliott *et al.* (2022) focuses on p-values under 0.15 and uses up to 60 bins. Since my dataset is much smaller, I use ten bins to focus on all p-values below 0.2. Results can be seen in Table 4.5. The p-value is above 0.1 (0.107 and 0.113) for both tests. Therefore, we do not reject the null hypothesis that p-hacking is absent. This test suggests there is no presence of p-hacking in my dataset, even at a 10% significance level. However, since results are only slightly above a 10% significance level, one cannot give that much weight to this one test. Another problematic aspect is that these tests require a larger sample size to be credible.

Table 4.4: MAIVE

MAIVE	
Coefficient	-0.017
(Standard error)	(0.480)
Observations	131
F-test	0.340

Note: This table presents the results of the MAIVE by Irsova *et al.* (2023a). F-test = Test statistic indicating strength of chosen instrument in first stage of the estimation. Cluster-robust standard errors are utilized in the MAIVE estimation. ***p<0.01, **p<0.05, *p<0.1

Table 4.5: p-hacking tests by Elliott

	Test for non-increasingness	**Test for monotonicity and bounds**
p-value	0.107	0.113
Observations ($p \leq 0.2$)	55	55
Total observations	131	131

Note: The table reports the conclusion of p-hacking tests by Elliott *et al.* (2022). It includes the non-increasigness test and the monotonicity and bounds test.

Another method of investigating publication bias is the Robust Bayesian Model Averaging (RoBMA) by Maier *et al.* (2022). It involves estimating several models and integrating these models through BMA. Every single model is given a specific weight. Afterward, Bayes factors are used to test for the presence of publication bias or heterogeneity individually.

See Table 4.6, for the results of RoBMA, I used the following specification: priors effect with Cauchy distribution (location = 0, scale = $1/\sqrt{2}$) and prior heterogeneity with inverse-gamma distribution (shape = 1, scale = 0.15). The specification is based on Cala (2024). RoBMA found strong evidence against the existence of teacher experience effect; it detected strong evidence in favor of heterogeneity and publication bias. RoBMA discovered a linear relation between effect sizes and standard errors. As mentioned earlier, it suggests zero effect of teacher experience; there were 36 models used to estimate the presence of the effect, heterogeneity, and publication bias. 18/36 models assumed the presence

of effect and heterogeneity, and 32/36 concluded the presence of publication bias.

Table 4.6: RoBMA — Robust Bayesian Model Averaging by Maier

	Model-averaged estimates			
	Mean	Median	0.025	0.975
Coefficient	0.000	0.000	0.000	0.000
Observations	131	131	131	131
	Individual effects			
	Models	Prior Prob.	Post. Prob.	Inclusion BF
Effect	18/36	0.500	0.000	0
Heterogeneity	18/36	0.500	1.000	∞
Bias	32/36	0.500	0.999	740

Note: The table presents the Robust Bayesian Model Averaging method results by Maier *et al.* (2022). The first part consists of statistics of the estimates produced by the model averaging method. The other part outlines the effect, heterogeneity, and publication bias. Models = Fraction of models that contain effect, heterogeneity, or publication bias. Prior Prob. = Prior Probability. Post. Prob. = Posterior Probability. Inclusion BF = Inclusion Bayes Factor. ***p<0.01, **p<0.05, *p<0.1

4.7 Summary of Publication Bias Effect

In previous chapters, the presence of publication bias was tested using various methods and models. The hypothesis was that there is a possibility of publication bias, or as it is sometimes named, p-hacking since there is a general consensus that teacher experience should positively influence test scores. For example, in the paper of Podolsky *et al.* (2019).

Firstly, I utilized Funnel Plot (Egger *et al.* 1997), which, in my eyes, showed slight signs of publication bias. Afterward, using linear tests, I found evidence of publication bias in case five out of six models. In the case of nonlinear tests, the Endogenous kink model from Bom & Rachinger (2019) suggested a strong presence of publication bias alongside the Selection model (Andrews & Kasy 2019), which indicated only a weak effect of publication bias.

Endogeneity-robust tests suggested conflicting results. Instrumental Vari-

able Regression suggested the presence of publication bias, while p-uniform* developed by Van Aert & Van Assen (2021).

P-hacking tests by Elliott *et al.* (2022) showed no sign of publication bias. On the other hand, RoBMA by Maier *et al.* (2022) suggested a high chance of publication bias presence.

These findings highlight the complexity and variability in detecting publication bias/p-hacking, demonstrating the need for a robust approach in analyzing potential publication bias. The mixed results from different tests and models underscore the importance of using a combination of methods to obtain a more comprehensive understanding of publication bias in educational research. The presence of p-hacking is not confirmed with absolute certainty; however, it is highly expected despite some tests yielding contrary results. The presence of publication bias was anticipated, given the widely accepted belief that teacher quality increases with experience. This belief might pressure researchers to publish results that align with this intuition, potentially skewing the available literature toward positive findings. Consensus on the positive impact of teacher experience is highlighted by the findings of Podolsky *et al.* (2019), Hanushek & Rivkin (2006), and Harris & Sass (2011).

Chapter 5

Heterogeneity

In this section, we investigate the significant variations in the estimates reported in the literature, as illustrated in Table 3.2 and Table 3.3. While publication bias partially explains this variation, individual studies employ diverse methods and samples, contributing to the observed discrepancies. The correlation between effect estimates and standard errors noted in previous analyses may stem both from publication bias and heterogeneity among studies. This section aims to ensure the validity of previous conclusions regarding publication bias and the true underlying effect by explicitly accounting for heterogeneity. Assessing the robustness of previous findings regarding publication bias is essential, as heterogeneity among study outcomes may lead to asymmetry in funnel plots, even in the absence of selective reporting.

In earlier sections, estimates of the teacher experience effect were obtained in various contexts without explicitly addressing this heterogeneity. Although some tests for publication bias allowed for systematic heterogeneity, none provided a comprehensive treatment. This section seeks to fill this gap with three primary objectives: assessing the robustness of publication bias findings when explicitly controlling for heterogeneity, identifying study arrangement characteristics that systematically affect results, and estimating the effect of teacher experience on student test score across different contexts after correcting for biases.

5.1 Study Context Variables

I have compiled 28 variables representing differences between studies to support this analysis. For clarity, these variables are categorized into three main groups:

Data Characteristics, Estimation Technique, and Publication Characteristics. Instances of Data Characteristics variables include the subject of the observed test scores (such as mathematics) and the type of student sample (for example, the grade students are in). The Estimation Technique category includes the methods used or control variables included in the regression of the reported estimate. Publication Characteristics encompass variables such as sample size and the number of citations. The mean, weighted mean, and standard deviation of these regression variables and their descriptions are presented in Table 5.1. Only relevant, quantifiable variables are included. Several variables serve as reference categories and are not further included in the regression analysis to avoid the dummy variable trap. Overall, I collected variables that were both widely available across many papers and appeared to have the potential to influence the teacher experience effect.

Data Characteristics

Interestingly, over 70% of the studies in the dataset were conducted in the USA, where the reported estimates were significantly higher. It necessitated the inclusion of a dummy variable for the country, as the USA's unique education system could lead to distinct effects of teacher experience. Additionally, it seemed logical to code dummy variables to categorize the studies based on the grade level of the students examined, as younger students might be influenced differently by their teachers compared to older students. The studies also varied by the subject analyzed, so I coded for whether the investigation focused on mathematics, reading, or other subjects.

Estimation methods

As discussed previously, studies utilize varying methods and model specifications, which may cause substantial systematical differences in reported estimates as they all have different assumptions and specific influences. I selected the most impactful ones; therefore, I coded four different methods and ten control variables.

The first control variable I coded was the inclusion of prior achievement of student in regression; it might be the most important one as it mitigates omitted variable bias in regression (Hanushek & Rivkin 2006). Including student prior achievement in regression is advantageous, as it is a highly significant variable with a crucial impact on future student performance due to its solid

predictive value (Getenet 2023). I also coded controlling for teacher education as it is intuitive education might influence the effect of teacher experience. Although I initially considered studying teacher education effect instead of teacher experience, the limited number of available estimates discouraged this approach. I coded whether regression included the control variable for teacher gender and ethnicity, respectively. Dee (2005) suggests students might be affected by teacher gender or ethnicity. According to Royer & Walles (2007), student gender and ethnicity influence test score. Therefore, I coded whether regression controls for student ethnicity and gender. The last three control variables, parent education, student ability, and peers' ability, had the same function, as they all reduce omitted variable bias in regression. Parent education is a proxy variable for economic background and student ability alongside peers' ability control for the fact more experienced teachers may move to teach in better schools (Hanushek *et al.* 2004).

I also coded for the model used to obtain the reported estimate. The most common was the Fixed Effects model, mainly in the latest papers; the second most frequent was Ordinary Least Squares; the least frequent was hierarchical linear modeling and random effects model. According to Podolsky *et al.* (2019), the fixed effects model allowed for the most accurate estimates compared to other models.

Publication characteristics

I collected variables regarding publication information. For every study, I coded a number of citations and publication year. Citations (Google Scholar was used as a source) were transformed using a natural logarithm. These variables are essential since the first is a proxy variable for the 'quality' of the paper, and the second enables capturing any changes throughout time. The study included in the analysis with the highest citation score was Hill *et al.* (2005), and the newest paper from Sancassani (2023) was published last year. Overall, almost all studies had a high number of citations.

5.2 Model Averaging

My objective is to measure how different factors influence the effect of teacher experience using Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA). BMA and FMA will aid in identifying the impact of varying

study specifications on the observed effects and assign a number to the extent of this influence.

In the subsequent chapter, I will build upon the results obtained using these methods to derive a best-practice estimate for both my analysis and the overall dataset. This approach will offer a much better and deeper understanding of the teacher experience effect on the student test score, incorporating the complexities and variations among different study contexts.

5.3 Bayesian model averaging

I collected 28 variables that could influence the effect of teacher experience on student test scores. However, only some of these variables will prove to be statistically significant in reality. This situation underscores the problem of model uncertainty, as it is not immediately clear which explanatory variables should be included in the model. With 2^{28} possible models, I need an effective way to determine which variables to use. Including all variables is not feasible, as it would make the estimates of significant variables inefficient and lead to over-specification bias. According to Steel (2020), addressing model uncertainty can be effectively achieved by adopting Bayesian model averaging (BMA) as a natural response.

The Bayesian model averaging technique addresses this issue by allowing me to estimate the probability that a particular explanatory variable should be included in the model. By utilizing Bayes' theorem and posterior inclusion probabilities, BMA assigns weights to multiple models based on the fit of each model. These weights then determine the importance of each factor. BMAproduces three key statistics for each factor: Posterior Mean, Posterior Standard Deviation, and Posterior Inclusion Probability (PIP). For the calculation of Posterior Inclusion Probability, it is crucial to understand Posterior Model Probability (PMP), which indicates the fit quality of each model. PIP of a variable is equal to the total PMP of all models that contain that particular variable. The higher the PIP, the more effectively the variable explains the differences in reported estimates of the teacher experience effect among papers (Hoeting *et al.* 1999).

For the specification of BMA implemented in this meta-analysis, the default Zellner's g-prior 'UIP' (regression coefficients, g=number of observations g=number of observations) was used, which assigns each coefficient the same weight as one data point, along with a dilution model prior. The choice of the

dilution model prior, developed by George (2010), stems from the idea that a relatively high number of variables are used, and collinearity might be an issue. Therefore, it is beneficial to use a prior that gives less weight to models featuring substantial collinearity. Additionally, a default sampler using the Markov Chain Monte Carlo (MCMC) algorithm class was employed (Zeugner & Feldkircher 2015). This methodological approach ensures that the final model is both parsimonious and robust, effectively mitigating the risks associated with over-specification bias and inefficient estimates.

5.4 Frequentist model averaging

Following the example of Cala (2024), as a further robustness check, I also include Frequentist Model Averaging using Mallow's criteria for weights (Hansen 2007) and orthogonalization of the model space as suggested by Amini & Parmeter (2012). The rationale behind using FMA includes its enhanced resilience to model misspecification and ability to reduce model uncertainty. FMA serves as an effective method to ensure that the Bayesian Model Averaging framework is correctly specified.

5.5 Results

Figure 5.1 presents graphical results of BMA. The figure's columns correspond to individual regression models, whereas the rows show which variables are included in each model. Each variable's impact is represented by blue for a positive effect and orange for a negative effect. White cells are a representation that variable was not included in the particular model. On the vertical axis, the explanatory variables are arranged according to their posterior inclusion probabilities, with the variables having the highest probabilities positioned at the top and those with the lowest probabilities at the bottom. This ranking highlights that the first variables mentioned are those with highest power in explaining the variations in the reported estimates of the teacher experience effect. The width of the columns represents the individual model's Posterior Model Probability. Models on the left side show the most optimal blend of data fit and resources used. PIP is a percentage representation of the factor in models. If the PIP of the variable is equal to 0.75, it is part of 75% of models.

In Table 5.2 and Table 5.3, we can see the numerical results of BMA and

FMA. Only five variables have PIP over 0.5, and including Standard Error, only three had PIP over 0.7. Generally, PIP results under 0.7 indicate a weaker effect, while results above 0.7 indicate a moderate effect of the factor. Factors included in the model with the best fit and at the same time with the highest PIP are standard error, controlling for teacher ethnicity, using the method of fixed effects, sampling of students from 9th to 12th grade, and controlling for student gender.

For FMA, not many variables have p-values low enough for them to be significant; however, few of them do, and they mostly correlate with those with high PIP in BMA. These include standard error, the method of fixed effects, controlling for teacher ethnicity, controlling for student gender, and barely above 10% significance with a p-value of 0.102 for a sample of students from grades 9th to 12th.

Looking at what these models can tell about the presence of publication bias, we can notice the effect of the standard error is present in both of them, with a PIP value of 1 and a p-value of 0.010. The standard error coefficient in the case of BMA is 0.281, while for FMA, it is 0.201. Even though the results of tests in Chapter 4 were not very conclusive, both BMA and FMA provide strong evidence that publication bias is present.

Among the variables with a significant impact (PIP over 0.5), only one shows a negative influence: the sample of 9th to 12th-grade students, with an average impact of -0.034 on the teacher experience effect. It might indicate that the teacher experience effect grows smaller for older students. As previously mentioned, the standard error positively impacts the effect, indicating publication bias. When regressions control for teacher ethnicity, the teacher experience effect increases by an average of 0.048. Similarly, controlling for student gender results in an average effect increase of 0.018, and using the fixed effects method increases the effect by 0.027 on average. It implies a possible downward bias that only the fixed effects model corrects. Other variables have a PIP below 0.5, rendering them less impactful, and their posterior means are typically minimal and close to zero.

In the context of Frequentist Model Averaging, the coefficients are as follows: 0.068 if regression was controlling for teacher ethnicity, 0.043 if regression was controlling for student gender, and 0.044 if reported estimates were obtained using the fixed effects method. Like BMA, the teacher experience effect is smaller for student samples from 9th to 12th grade, with a coefficient of -0.050.

In Figure B.1 and Figure B.2, robustness checks are included using different

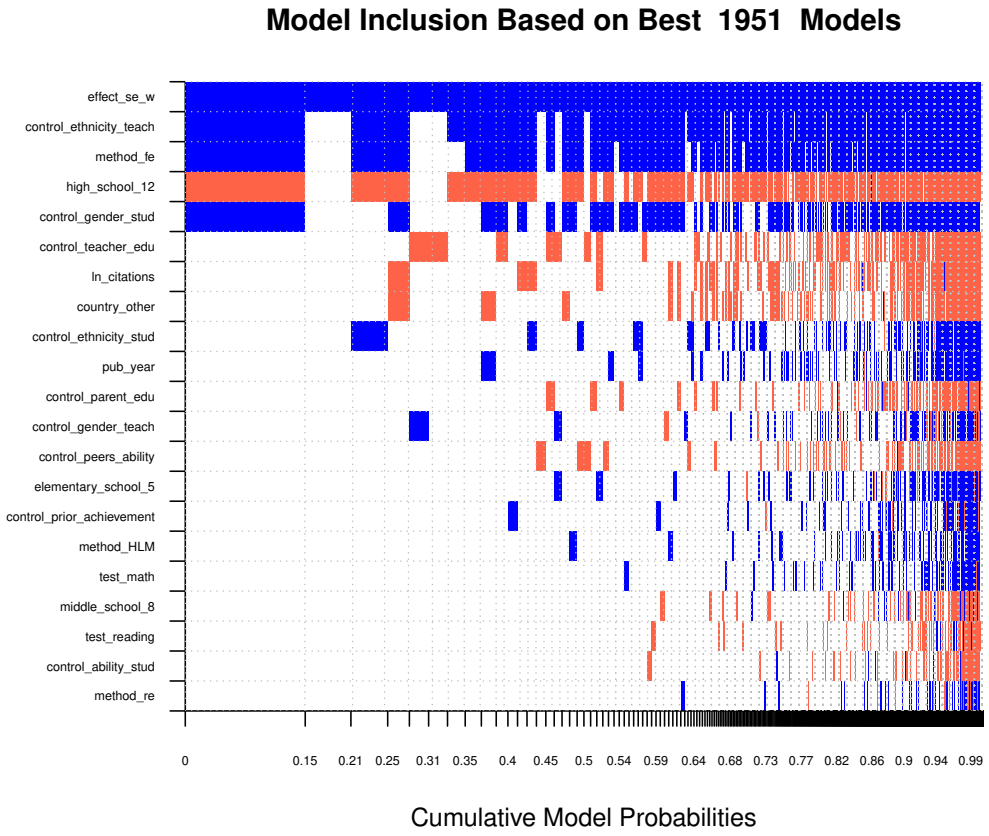
specifications of BMA, specifically a uniform g-prior with a uniform model prior and an HQ g-prior with a random model prior. The results from these checks are not surprising, as they are very similar to those described previously. This consistency further validates the robustness of the findings obtained using the BMA method.

Table 5.1: Summary of variables used in heterogeneity analysis and their meaning

Variable	Description	Mean	SD	Weighted Mean
Effect	Reported estimate of the effect of one year of teacher experience on test score (one standard deviation change)	0.0198	0.0700	0.0235
Standard error	Standard error of the estimate	0.0505	0.1590	0.0503
Data Characteristics				
General population	=1 if sample consisted of non-specific student sample	0.9847	0.1231	0.9850
Study USA	=1 if study was conducted in the USA (reference)	0.7023	0.4590	0.7894
Study Other country	=1 if study was not conducted in the USA	0.2977	0.4590	0.2106
Math test	=1 if math test was used	0.4733	0.5012	0.5641
Reading test	=1 if reading test was used	0.1985	0.4004	0.1682
Other subject test	=1 if test was neither math or reading (reference)	0.3206	0.4685	0.2611
Kindergarten to 5th grade	=1 if sample consisted of students from kindergarten to 5th grade	0.5115	0.5018	0.4471
6th to 8th grade	=1 if if sample consisted of students from 6th to 8th grade	0.2214	0.4168	0.2106
9th to 12th grade	=1 if sample consisted of students from 9th to 12th grade	0.2137	0.4115	0.2896
Different grade	=1 if sample consisted of students from different mix of grades (reference)	0.0534	0.2258	0.0527
Disadvantaged students	=1 if sample consisted of disadvantaged students: (1) eligible for free or reduced price lunch 2) students with achievement scores in bottom 25 percentile 3) minority	0.0229	0.1502	0.0226
Advantaged students	=1 if sample consisted of advantaged students: (1) academically gifted 2) top 25 percentile of achievement	0.0076	0.0874	0.0075
Estimation Technique				
Control: Prior achievement	=1 if authors control for prior achievement of student in the regression	0.6336	0.4837	0.6138
Control: Student Gender	=1 if authors control for student gender in the regression	0.6260	0.4857	0.5562
Control: Student Ethnicity	=1 if authors control for student ethnicity in the regression	0.5954	0.4927	0.5436
Control: Parent Education	=1 if authors control for parent education in the regression	0.2595	0.4401	0.3335
Control: Teacher Education	=1 if authors control for teacher education in the regression	0.7710	0.4218	0.6269
Control: Student Ability	=1 if authors control for student ability in the regression	0.0611	0.2404	0.1053
Control: Peers' Ability	=1 if authors control for peers' ability in the regression	0.3511	0.4792	0.2633
Control: Teacher Gender	=1 if authors control for teacher gender in the regression	0.5191	0.5016	0.4560
Control: Teacher Ethnicity	=1 if authors control for teacher ethnicity in the regression	0.2824	0.4519	0.3507
Control: Student Age	=1 if authors control for student age in the regression	0.0076	0.0874	0.0527
Method: FE	=1 if fixed effects method was used	0.4351	0.4977	0.4349
Method: OLS	=1 if ordinary least squares method was used (reference)	0.3817	0.4877	0.3369
Method: HLM	=1 if hierarchical linear model was used	0.0916	0.2896	0.1229
Method: RE	=1 if random effects method was used	0.0916	0.2896	0.1053
Publication Characteristics				
Publication Year	Year when was the paper published	2008	8.1814	2008
ln_citations	Natural log of the number of citations in Google Scholar	5.6701	1.7731	5.6837

Note: The table provides information and description about various study characteristics. Reference variables are excluded from the Bayesian model averaging estimate (to avoid dummy variable trap). SD = standard deviation, WM = mean using the inverse amount of estimates included in each study as a weight

Figure 5.1: Graphical results of Bayesian model averaging



Note: The figure presents graphical results of the Bayesian model averaging. The following specification was used: the uniform g-prior and dilution prior model. The dependent variable, the test score increase (in standard deviation) per additional year of teacher experience, is estimated on the horizontal axis as cumulative posterior model probabilities. The independent variables are ranked (highest to lowest) on the vertical axis based on their PIP. The associated independent variable has a positive (negative) sign in the case of blue (orange) color. The associated independent variable is not part of the model in the case of white color. Table 5.2 provide numerical results of the estimation. Detailed description of the variables is included in Table 5.1

Table 5.2: Bayesian model averaging

Response variable:	Effect of one year of teacher experience		
	P. Mean	P. SD	PIP
(Constant)	-0.2860	NA	1.0000
Effect Standard Error	0.2805	0.0522	1.0000
Data Characteristics			
high_school_12	-0.0340	0.0256	0.6946
country_other	-0.0040	0.0101	0.1716
elementary_school_5	0.0011	0.0054	0.0747
test_math	0.0003	0.0020	0.0429
middle_school_8	-0.0004	0.0037	0.0426
test_reading	-0.0002	0.0018	0.0306
Estimation Technique			
control_ethnicity_teach	0.0484	0.0285	0.8076
method_fe	0.0269	0.0185	0.7478
control_gender_stud	0.0180	0.0184	0.5504
control_teacher_edu	-0.0047	0.0098	0.2298
control_ethnicity_stud	0.0039	0.0100	0.1651
country_other	-0.0040	0.0101	0.1716
control_parent_edu	-0.0017	0.0066	0.0918
control_gender_teach	0.0013	0.0058	0.0904
control_peers_ability	-0.0013	0.0053	0.0881
method_HLM	0.0006	0.0035	0.0495
test_math	0.0003	0.0020	0.0429
middle_school_8	-0.0004	0.0037	0.0426
control_ability_stud	-0.0002	0.0030	0.0295
method_re	0.0000	0.0016	0.0224
Publication Characteristics			
ln_citations	-0.0013	0.0029	0.1990
pub_year	0.0001	0.0005	0.1066

Note: The table provides the results of the Bayesian model averaging. P. mean = Posterior Mean, P. SD = Posterior Standard Deviation, PIP = Posterior Inclusion Probability. Detailed description of the variables is included in Table 5.1

Table 5.3: Frequentist model averaging

Response variable:	Effect of one year of teacher experience		
	Coefficient	Standard Error	p-value
(Constant)	-3.9312	2.6317	0.1350
Effect Standard Error	0.2006	0.0302	0.0000
Data Characteristics			
high_school_12	-0.0504	0.0308	0.1020
country_other	-0.0292	0.0218	0.1800
elementary_school_5	-0.0273	0.0287	0.3410
test_math	0.0140	0.0105	0.1820
middle_school_8	-0.0464	0.0333	0.1640
test_reading	0.0126	0.0125	0.3130
Estimation Technique			
control_ethnicity_teach	0.0680	0.0225	0.0030
method_fe	0.0444	0.0137	0.0010
control_gender_stud	0.0434	0.0188	0.0210
control_teacher_edu	-0.0044	0.0116	0.7040
control_ethnicity_stud	0.0017	0.0246	0.9450
country_other	-0.0292	0.0218	0.1800
control_parent_edu	-0.0235	0.0171	0.1690
control_gender_teach	0.0141	0.0190	0.4580
control_peers_ability	-0.0040	0.0114	0.7260
method_HLM	0.0142	0.0112	0.2050
test_math	0.0140	0.0105	0.1820
middle_school_8	-0.0464	0.0333	0.1640
control_ability_stud	-0.0184	0.0219	0.4010
method_re	-0.0049	0.0146	0.7370
Publication Characteristics			
ln_citations	-0.0069	0.0052	0.1850
pub_year	0.0020	0.0013	0.1240

Note: The table provides the results of the Frequentist model averaging. Detailed description of the variables is included in Table 5.1

Chapter 6

Best-practice estimate

In previous chapters, no definitive value has been established as the true effect of teacher experience on student test score standard deviations after accounting for publication bias and heterogeneity. Therefore, in this chapter, I aim to develop a best-practice estimate of the effect of teacher experience on student test scores by utilizing Bayesian Model Averaging model coefficients obtained in Chapter 5 and actual values from my dataset as is suggested by Irsova *et al.* (2023b). Since the concept of best practice can be subjective, I will adjust certain variables based on my judgment to approximate the true effect. I will report both my subjective estimate and an estimate based on all 19 studies included in my analysis.

To derive these estimates, I retrieved the BMA coefficients obtained in the previous chapter and multiplied them by the actual values from my dataset. However, I will predominantly use their sample mean values because best practice cannot be definitively determined for many of the variables. For instance, no clear consensus exists on which control variables should be included or which subjects should be used to test student achievement. Consequently, my subjective 'best-practice' estimate will involve two critical adjustments: first, I set the standard error to zero to mitigate the impact of publication bias. Second, I used the maximum value of the number of citations (in natural logarithm form) under the assumption that the most cited papers might be closer to reflecting the true effect. For the remaining variables, I used their sample mean values.

'Best-practice' estimate of all 19 studies. also involved replacing the standard error's value with zero to correct the estimates for the effect of publication bias. Since I used all the values in the 19 studies for the calculation, I had to

address instances where the same paper had provided multiple values of a variable, which I solved by using their sample means.

The results are presented in Table 6.1. As shown, there is a relatively sizable difference in estimates. The estimate based on all 19 studies suggests an increase of 0.0057 in test score standard deviation per additional year of teaching experience, which is almost twice as large as my subjective estimate of 0.0023. The 95% confidence intervals are vast for both estimates and partially overlap with negative values. It indicates that after correcting for publication bias and accounting for heterogeneity, the estimate size decreases, and it is not clear whether the effect is even different from zero. The wide confidence intervals likely result from aggregating many papers with varying specifications into a single estimate or from the lack of consensus in the literature on the magnitude of this effect. Although one might expect teacher experience to impact student achievement significantly, best-practice estimates suggest either a negligible effect or no effect based on the papers included in this meta-analysis.

Table 6.1: 'Best-practice' estimate

Study	**Estimate**	**Standard Error**	**95% Confidence Interval**	**Studies**
Author	0.0023	0.0166	(-0.0302; 0.0348)	0
All studies	0.0057	0.0101	(-0.0141; 0.0256)	19

Note: This table presents best-practice estimates based on the author's subjective knowledge and the complete dataset. The results are calculated by averaging the best-practice estimates of every single study. Confidence interval at 95% level is calculated by utilizing OLS using standard errors clustered at the study level.

Chapter 7

Conclusion

I conducted a comprehensive analysis to study the relationship between teacher quality, measured by teacher experience, and student achievement, measured by standardized test scores. I gathered 131 effect estimates from 19 studies, including various methodological and sample characteristics. Initially, I calculated the average effect reported in the papers, which indicated a 0.02 increase in student test score standard deviation per one additional year of teacher experience. I then used advanced methods to test publication bias and calculate effect corrected for this bias.

Firstly, I utilized a simple graphical test developed by Egger *et al.* (1997) called the funnel plot. The following tests can be sorted into three categories: linear methods - Funnel Asymmetry Tests - Precision Effect Testing with different specifications. Non-linear methods, which do not rely on the linear structure of the effect and standard error relationship, are represented by the Stem-based method (Furukawa 2019), the Weighted Average of the Adequately Powered (Ioannidis *et al.* 2017), the Top10 method (Stanley *et al.* 2010), the Endogenous kink model (Bom & Rachinger 2019), and the Selection model (Andrews & Kasy 2019). Endogeneity-robust methods - the p-uniform* method developed by Van Aert & Van Assen (2021), Elliot tests (Elliott *et al.* 2022), and the MAIVE estimator recently developed by Irsova *et al.* (2023a). The presence of publication bias is indicated by five out of six linear tests, suggested by four non-linear methods, and assumed by two out of four latest techniques. Overall, results indicate the presence of publication bias, and when correcting for it, the effect size decreases considerably, moving closer to zero.

I utilized Bayesian (Maier *et al.* 2022) and Frequentist model averaging approaches (Steel 2020) to examine heterogeneity. I identified five variables

that have a significant impact on the effect size. These were standard error, a sample of students between 9th and 12th grade, use of the fixed effects method, controlling for the student's gender in regression, and controlling for teacher ethnicity, only the 9th to 12th grade sample had a negative influence on effect size. The rest of the variables had small and not very significant impacts. To further inspect the model results, I also included a robustness check of Bayesian model averaging using different specifications, which can be found in the Appendix.

In the last chapter, I calculated the best-practice estimate, which aligns with the results of previously conducted tests. The effect is not significantly different from zero when the specification is set to obtain a number closest to the true effect and when accounting for publication bias. Overall, the results are inconclusive as some tests returned tiny negative results, some were similar to zero, and several produced slightly positive estimates. This indicates either a lack of evidence in identified papers or that the effect does not exist. Insignificant effect contradicts the results in current literature, for example, represented by a review of Podolsky *et al.* (2019), which concluded that teaching experience positively affects student achievement gains. There is a possibility of replicating this analysis in the future with more papers published on this topic, which might lead to different conclusions, as one of the downsides of this analysis is the seemingly low number of papers and estimates, as many papers on this topic did not include all the necessary data. There needs to be further research conducted in this field to observe other teacher variables that contribute more in development of teacher quality, because teacher remains one of the most influential factors in education as stated by Hanushek (2011).

Bibliography

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

Anderson, T. W. & H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations." *The Annals of Mathematical Statistics* **20(1)**: pp. 46–63.

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

Betts, J. R. & J. L. Shkolnik (2000): "The effects of ability grouping on student achievement andresource allocation in secondary schools." *Economics of Education Review* **19**: pp. 1–15.

Blazar, D. (2015): "Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement." *Economics of Education Review* **48**: pp. 16–29.

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

Borman, G. D. & S. M. Kimball (2005): "Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps?" *The Elementary School Journal* **106(1)**.

Cala, P. (2024): "Ability bias in the returns to schooling: How large it is and why it matters."

Canales, A. & L. Maldonado (2018): "Teacher quality and student achievement in chile: Linking teachers'contribution and observable characteristics." *International Journal of Educational Development* **60**: pp. 33–50.

Clotfelter, C. T., H. F. Ladd, & J. L. Vigdor (2010): "Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects." *Journal of Human Resources* **45(3)**: pp. 655–681.

Cremata, E., D. Davis, K. Dickey, K. Lawyer, Y. Negassi, M. E. Raymond, & J. L. Woodworth (2013): "National charter school study 2013." *Technical report*, Stanford University, Stanford, CA.

Darling-Hammond, L. (1999): "Teacher quality and student achievement: A review of state policy evidence." *Education Policy Analysis Archive* **8(1)**: pp. 1–44.

Darling-Hammond, L., D. J. Holtzman, S. J. Gatlin, & J. V. Heilig (2005): "Does teacher preparation matter? evidence about teacher certification, teach for america,and teacher effectiveness." *Education Policy Analysis Archives* **13(42)**.

De Paola, M. (2009): "Does teacher quality affect student performance? evidence from an italian." *Bulletin of Economic Research* **61(4)**: pp. 353–377.

Dee, T. S. (2005): "A teacher like me: Does race, ethnicity, or gender matter?" *American Economic Review* **95(2)**: pp. 158–165.

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.

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

Furukawa, C. (2019): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." Working paper, Massachusetts Institute of Technology.

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." **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.

Getenet, S. (2023): "The influence of students' prior numeracy achievement on later numeracy achievement as a function of gender and year levels." *Mathematics Education Research Journal* .

Glass, G. V. & M. L. Smith (1979): "Meta-analysis of research on class size and achievement." *Educational Evaluation and Policy Analysis* **1(1)**: pp. 2–16.

Goldhaber, D. D. & D. J. Brewer (1996): "Evaluating the effect of teacher degree level on educational performance." .

Goldhaber, D. D. & D. J. Brewer (2000): "Does teacher certification matter? high school teacher certification status and student achievemen." *Educational Evaluation and Policy Analysis* **22(2)**.

Graham, L. J., S. L. J. White, K. Cologon, & R. C. Pianta (2020): "Do teachers' years of experience make a difference in the quality of teaching?" *Teaching and Teacher Education* **96**.

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

Hanushek, E. A. (2011): "The economic value of higher teacher quality." *Economics of Education Review* **30(3)**: pp. 466–479.

Hanushek, E. A., J. F. Kain, & S. G. Rivkin (2004): "Why public schools lose teachers." *Journal of Human Resources* **39(2)**: pp. 326–354.

Hanushek, E. A. & S. G. Rivkin (2006): "Teacher quality." In "Handbook of the Economics of Education," pp. 1051–1078.

Harris, D. N. & T. R. Sass (2011): "Teacher training, teacher quality and student achievement." *Journal of Public Economics* **95(7-8)**: pp. 798–812.

Hassan, A., N. S. Jamaludin, T. Sulaiman, & R. Baki (2010): "Western and eastern educational philosophies." In "40th Philosophy of Education Society of Australasia Conference," .

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2022): "Skilled and unskilled labor are less substitutable than commonly thought." *The Review of Economics and Statistics* .

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2024): "Publication and attenuation biases in measuring skill substitution." *The Review of Economics and Statistics* .

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

Hill, H. C., B. Rowan, & D. L. Ball (2005): "Effects of teachers' mathematical knowledge for teaching on student achievement." *American Educational Research Journal* **42(2)**: pp. 371–406.

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

Ioannidis, J. P., T. D. Stanley, & H. Doucouliagos (2017): "The power of bias in economics research." *Economic Journal* **127(605)**: pp. 236–265.

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

Irsova, Z., H. Doucouliagos, T. Havranek, & T. D. Stanley (2023b): "Meta-Analysis of Social Science Research: A Practitioners Guide." **(273719)**.

Jepsen, C. (2005): "Teacher characteristics and student achievement: evidence from teacher surveys." *Journal of Urban Economics* pp. 302–319.

Kingdon, G. & F. Teal (2010): "Teacher unions, teacher pay and student performance in india: A pupil fixed effects approach." *Journal of Development Economics* **91**: pp. 278–288.

Krieg, J. M. (2006): "Teacher quality and attrition." *Economics of Education Review* **25**: pp. 13–27.

Kukla-Acevedo, S. (2008): "Do teacher characteristics matter? new results on the effects ofteacher preparation on student achievement." *Economics of Education Review* **28**: pp. 49–57.

Leigh, A. (2010): "Estimating teacher effectiveness from two-year changes in students' test scores." *Economics of Education Review* **29**: pp. 480–488.

Maier, M., F. Bartos, & E.-J. Wagenmakers (2022): "Robust bayesian meta-analysis: Addressing publication bias with model-averaging." *Psychological Methods* **28(1)**: pp. 107–122.

Miller, R. T., R. J. Murnane, & J. B. Willett (2008): "Do teacher absences impact student achievement? longitudinal evidence from one urban school district." *Educational Evaluation and Policy Analysis* **30(2)**: pp. 181–200.

Munoz, M. A. & F. C. Chang (2007): "The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change." *Journal of Personnel Evaluation in Education* **20(3-4)**: pp. 147–164.

Opatrny, M., T. Havranek, Z. Irsova, & M. Scasny (2023): "Class size and student achievement: A modern meta-analysis." .

Papay, J. P. & M. A. Kraft (2015): "Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement." *Journal of Public Economics* **130**: pp. 105–119.

Penner, E. K. (2021): "Teach for americaand teacher quality: Increasing achievement over time." *Educational Policy* **35(7)**: pp. 1047–1084.

Pham, L. D., T. D. Nguyen, & M. G. Springer (2021): "Teacher merit pay: A meta-analysis." *American Educational Research Journal* **58(3)**: pp. 527–566.

Podolsky, A., T. Kini, & L. Darling-Hammond (2019): "Does teaching experience increase teacher effectiveness? a review of us research." *Journal of Professional Capital and Community* **4(4)**: pp. 286–308.

Reeves, P. M., W. H. Pun, & K. S. Chung (2016): "Influence of teacher collaboration on job satisfaction and student achievement." *Teaching and Teacher Education* **67**: pp. 227–236.

Rose, H. (2006): "Do gains in test scores explain labor market outcomes?" *Economics of Education Review* **25(4)**: pp. 430–446. Available online at www.sciencedirect.com.

Royer, J. M. & R. Walles (2007): "Influences of gender, ethnicity, and motivation on mathematical performance." In D. B. Berch & M. M. M. Mazzocco (editors), "Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities," pp. 349–367. Paul H. Brookes Publishing Co.

Sancassani, P. (2023): "The effect of teacher characteristics on students' science achievement." *Leibniz Institute for Economic Research at the University of Munich* .

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

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., S. B. Jarrell, & H. Doucouliagos (2010): "Could it be better to discard 90% of the data? a statistical paradox." *Journal of Economic Literature* .

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

Sutton, A. & I. Soderstrom (1999): "Predicting elementary and secondary school achievement with school-related and demographic factors." *The Journal ofEducational Research* **92(6)**: pp. 330–338.

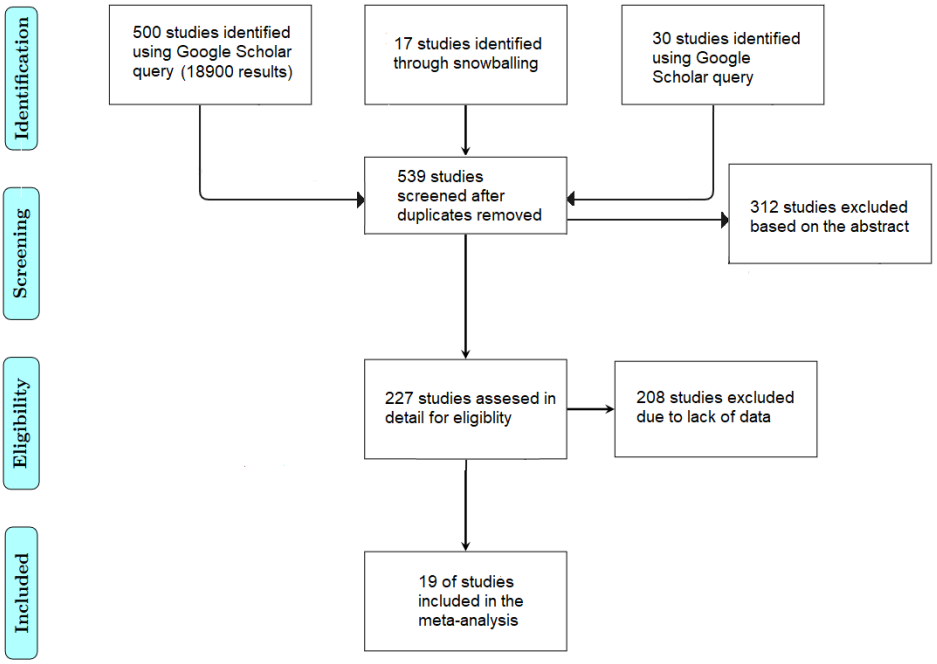
Van Aert, R. C. & M. Van Assen (2021): "Correcting for publication bias in a meta-analysis with the p-uniform* method." Working paper, Tilburg University & Utrecht University, available online at osf.io/preprints/metaarxiv/zqjr9/download.

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.

Appendix A

Literature Search Details

Figure A.1: Literature Search Details — PRISMA flow diagram

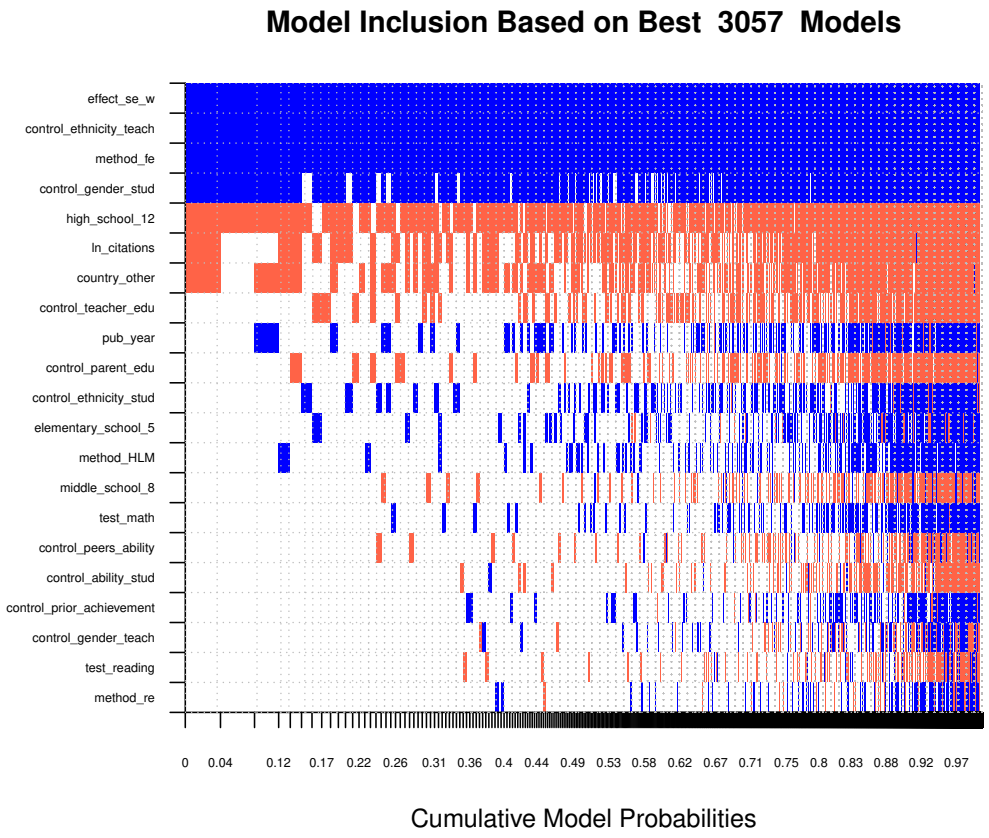


Note: The figure presents a PRISMA flow diagram illustrating the process of including studies in the meta-analysis. The following Google Scholar query was used in the search: ("teacher experience" OR "teacher quality" OR "teacher education" OR "teacher training") AND ("student achievement" OR "test scores" OR "educational outcomes"). The query search was conducted in April 2024. It searched five hundred studies and thirty studies from the last three years. It consisted of reading the abstract and inspecting the availability of data. The snowballing was conducted in May 2024. Inclusion criteria: 1. The study has to contain an estimated empirical relationship between teacher experience and test score. 2. The study has to contain standard errors or other information from which standard errors can be obtained. 3. The study must contain the test scores' standard deviation to convert estimates to a standardized form.

Appendix B

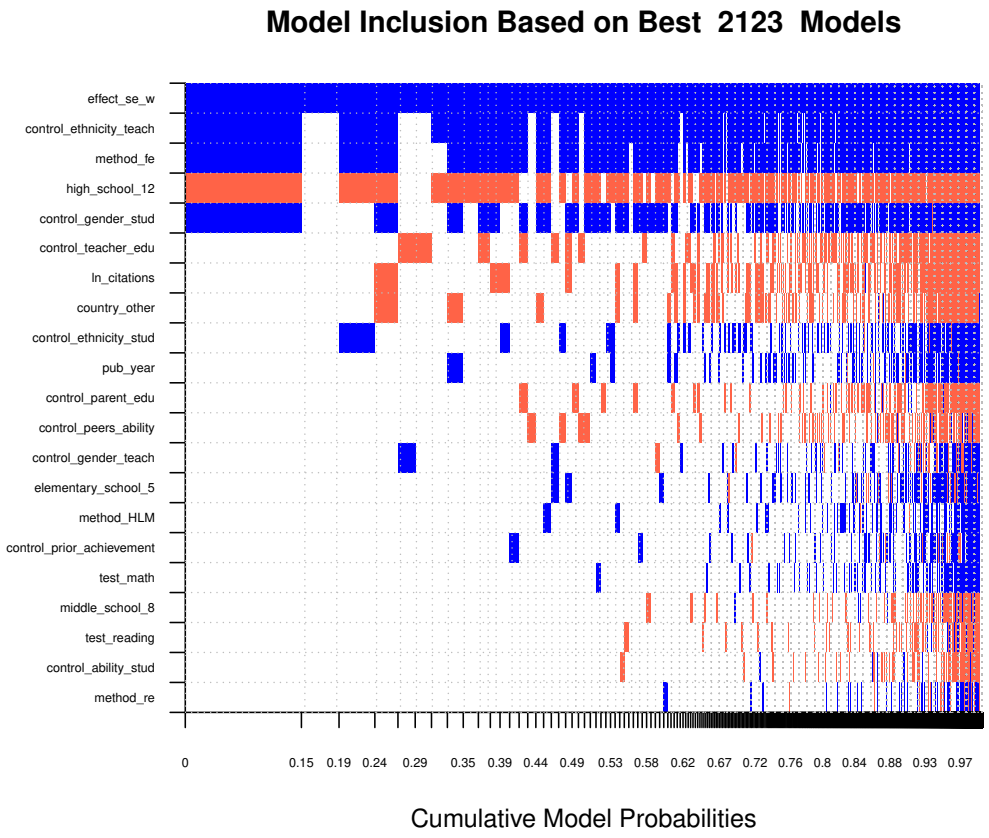
Bayesian model averaging robustness check

Figure B.1: Robustness check - BMA



Note: This graph presents the results of the Bayesian model averaging. It serves as a robustness check by applying different specifications compared to BMA results included in the main text of the thesis. Specifically, it utilizes the uniform g-prior alongside the uniform model prior.

Figure B.2: Robustness check - BMA



Note: This graph presents the results of the Bayesian model averaging. It serves as a robustness check by applying different specifications compared to BMA results included in the main text of the thesis. Specifically, it utilizes the HQ g-prior alongside the random model prior.

Appendix C

Implied teacher experience effect in literature

Table C.1: Implied mean effect ('best-practice') of teacher experience in literature

Study	**Estimate**	**Standard Error**	**95% Confidence Interval**
Author's estimate	0.0023	0.0166	(-0.0302; 0.0348)
Goldhaber & Brewer (1996)	0.0071	0.0036	(0.0000; 0.0142)
Sutton & Soderstrom (1999)	0.0088	0.0074	(-0.0057; 0.0234)
Goldhaber & Brewer (2000)	0.0062	0.0056	(-0.0047; 0.0171)
Betts & Shkolnik (2000)	-0.0037	0.0009	(-0.0056; -0.0019)
Hill et al. (2005)	0.0246	0.0210	(-0.0167; 0.0658)
Borman & Kimball (2005)	-0.0217	0.0207	(-0.0623; 0.0188)
Jepsen (2005)	0.0018	0.0038	(-0.0057; 0.0092)
Darling-Hammond et al. (2005)	-0.0069	0.0035	(-0.0137; -0.0002)
Krieg (2006)	-0.0031	0.0059	(-0.0147; 0.0085)
Munoz & Chang (2007)	-0.0117	0.0084	(-0.0282; 0.0048)
Miller et al. (2008)	0.0430	0.0167	(0.0102; 0.0758)
Kukla-Acevedo (2008)	0.0778	0.0301	(0.0189; 0.1367)
Kingdon & Teal (2010)	0.0111	0.0079	(-0.0045; 0.0266)
Leigh (2010)	-0.0007	0.0053	(-0.0110; 0.0096)
Blazar (2015)	0.0061	0.0121	(-0.0177; 0.0298)
Reeves et al. (2016)	-0.0098	0.0220	(-0.0529; 0.0333)
Canales & Maldonado (2018)	0.0049	0.0071	(-0.0090; 0.0188)
Penner (2021)	-0.0258	0.0082	(-0.0417; -0.0098)
Sancassani (2023)	0.0014	0.0020	(-0.0026; 0.0053)

Note: This table contains estimates of the 'best practice' for every single study. It also includes the 'best practice' created by the author. Confidence interval at 95% level is calculated by utilizing OLS using standard errors clustered at the study level.