CHARLES UNIVERSITY
Faculty of Social Sciences
Institute of Economic Studies

Female Leadership and ESG Performance: A Meta-Analysis

Master's thesis

Author: Bc. Karolína Hozová
Study program: Economics and Finance
Supervisor: prof. PhDr. Tomáš Havránek, Ph.D.
Year of defense: 2025

Declaration of Authorship

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

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

Prague, April 29, 2025

Karolína Hozová

Abstract

This thesis provides a quantitative overview of the existing literature that examines the impact of female leadership on firms' ESG performance. Using meta-analytic techniques, we test for the presence of publication selection bias through graphical and statistical methods and analyze the factors driving heterogeneity among reported estimates using Bayesian model averaging. Our analysis relies on a unique database comprising 547 estimates drawn from 108 primary studies. The results reveal evidence of positive publication selection bias in the literature, while the bias-corrected average effect of female leadership on ESG performance is negligible. Furthermore, we identify systematic heterogeneity in the reported estimates, with the most significant determinant being the geographical location of the analyzed firms. In particular, studies focusing on firms in the Middle East tend to report a systematically larger gender diversity effect, whereas different patterns emerge for studies examining Asian, US and European companies. Furthermore, controlling for the presence of a CSR committee is associated with a lower female leadership premium. Despite statistical significance, observed effects exhibit a very low economic magnitude, suggesting that while gender diversity in leadership may modestly contribute to ESG performance, it does not represent a dominant factor.

JEL Classification:	G34, M14, C11, C80
Keywords:	meta-analysis, publication bias, Bayesian Model Averaging, ESG score, female leadership, board gender diversity
Title:	Female Leadership and ESG Performance: A Meta-Analysis

Abstrakt

Táto diplomová práca poskytuje kvantitatívny prehľad o existujúcej literatúre skúmajúcej vplyv žien vo vedení na ESG výkonnosť firiem. Využitím metaanalytických techník testujeme prítomnosť publikačnej selektivity pomocou grafických a štatistických metód a analyzujeme faktory vysvetľujúce heterogenitu medzi reportovanými odhadmi využitím Bayesovského modelového priemerovania. Naša analýza je založená na unikátnej databáze obsahujúcej 547 odhadov zo 108 primárnych štúdií. Výsledky ukazujú, že v literatúre dochádza k pozitívnej publikačnej selektivite, pričom očistený priemerný efekt ženského vedenia na ESG výkonnosť je zanedbateľný. Zároveň nachádzame systematickú heterogenitu v odhadoch, pričom najvýznamnejším determinantnom je geografická poloha analyzovaných firiem. Konktrétne, štúdie zamerané na Blízky východ majú tendenciu systematicky vykazovať vyšší efekt rodovej diverzity, zatiaľ čo pri štúdiách analyzujúcich ázijské, americké a európske spoločnosti sa objavujú odlišné vzorce. Navyše, kontrola prítomnosti komisie pre spoločenskú zodpovednosť firiem je spojená s nižším efektom ženského vedenia. Napriek štatistickej významnosti však zistené efekty vykazujú veľmi nízku ekonomickú magnitúdu, čo naznačuje, že hoci rodová diverzita môže mierne prispievať k Environmental, Social and Governance (ESG) výkonnosti firiem, nepredstavuje dominantný faktor.

Klasifikácia JEL: G34, M14, C11, C80

Klúčové slová: meta-analýza, publikačná selektivita, Bayesovské modelové priemerovanie, ESG skóre, ženy vo vedení, rodová diverzita vo vedení

Názov práce: Ženy vo vedení a ESG výkonnosť firiem: Meta-analýza

Acknowledgments

The author would like to express her gratitude to her supervisor, prof. PhDr. Tomáš Havránek, Ph.D.for his time, patience, valuable suggestions and guidance throughout the process of writing this thesis.

This thesis is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 870245.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Hozová, Karolína: *Female Leadership and ESG Performance: A Meta-Analysis.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2025, pages 99. Advisor: prof. PhDr. Tomáš Havránek, Ph.D.

Contents

List of Tables **ix**

List of Figures **x**

Acronyms **xi**

Master's Thesis Proposal **xii**

1 Introduction **1**

2 Conceptual Framework **3**

2.1 Female Leadership Trends 3

2.2 Topics in ESG 5

2.3 ESG Ratings 6

2.4 Exploring ESG Through Leadership 8

2.5 Why Women Directors Matter for ESG 8

3 Notes on Estimation Techniques **10**

3.1 Estimation Methods 11

3.2 Measurement of Female Leadership 14

3.3 Measurement of ESG Performance 15

4 Data **17**

4.1 Identification of Primary Studies 17

4.2 Dataset Compilation 20

4.3 Descriptive Evidence 22

5 Unveiling Publication Bias **26**

5.1 Graphical Method 27

5.2 Linear Methods 28

5.3 Non-linear Methods 31

5.4 Caliper Test 33
5.5 Robustness Checks 34

6 Drivers of Heterogeneity **36**
6.1 Description of Effect Drivers 36
6.2 Estimation 40
6.3 Results 43
6.4 Subsample Robustness Check 48
6.5 The Best Practice Estimate 49

7 Conclusion **52**

Bibliography **55**

A Additional Information for Data Collection **I**

B Robustness Checks for Publication Bias **II**

C BMA Robustness Checks and Diagnostics **IV**

List of Tables

4.1 Primary studies used 19
4.2 Board gender diversity effects in different contexts 25

5.1 Linear tests 30
5.2 Non-Linear tests 31
5.3 Caliper test 34

6.1 Overview and descriptive statistics of contextual variables 39
6.2 Results of baseline BMA and frequentist check estimations 45
6.3 Predicted best practice estimates 50

B.1 Linear tests for a subsample of of preferred estimates III
B.2 Non-linear tests for a subsample of preferred estimates III
B.3 Caliper test for a subsample of preferred estimates III
C.1 Results of robustness BMA and FMA VI
C.2 Diagnostics of the baseline BMA (UIP and dilution priors) VII
C.3 Diagnostics of the alternative BMA (BRIC and random priors) . . VIII
C.4 Results of subsample robustness BMA and frequentist check estimations IX
C.5 Diagnostics of the subsample robustness BMA X

List of Figures

4.1 Distribution of gender diversity effects 22
4.2 Variation of estimates within and across countries 23

5.1 Funnel plot of reported estimates 28
5.2 Distribution of t-statistics of estimates 33

6.1 Model inclusion in Bayesian model averaging 44
6.2 Model inclusion in subsample robustness BMA 48

A.1 PRISMA flow diagram I
B.1 Funnel plot for a subsample of preferred estimates II
C.1 Correlations between potential predictors of heterogeneity IV
C.2 Model inclusion in robustness BMA V
C.3 Model size and convergence for the baseline BMA VII
C.4 Model size and convergence for the robustness BMA VIII
C.5 Model size and convergence for the subsample robustness BMA . X

Acronyms

BMA Bayesian Model Averaging
BRIC Bayesian Risk Information Criterion
CSR Corporate Social Responsibility
CSRD Corporate Sustainability Reporting Directive
ESG Environmental, Social and Governance
ESGD Environmental, Social and Governance Disclosure
FAT Funnel Asymmetry Test
FE Fixed Effects
FMA Frequentist Model Averaging
GLS Generalized Least Squares
GMM Generalized Method of Moments
GRI Global Reporting Initiative
ISS Institutional Shareholder Services
JCR Journal Citation Reports
KLD Kinder, Lydenberg, and Domini
LIML Limited Information Maximum Likelihood
LSEG London Stock Exchange Group
MSCI Morgan Stanley Capital International
NFRD Non-Financial Reporting Directive
NYSE New York Stock Exchange
OLS Ordinary Least Squares
PET Precision Estimate Test
PIP Posterior Inclusion Probability
PMP Posterior Model Probability
RE Random Effects
RIF Recursive Impact Factor
SASB Sustainability Accounting Standards Board
SEC Securities and Exchange Commission
2SLS Two-Stage Least Squares
UIP Unit Information Prior
WAAP Weighted Average of Adequately Powered

Master's Thesis Proposal

Author	Bc. Karolína Hozová
Supervisor	prof. PhDr. Tomáš Havránek, Ph.D.
Proposed topic	Female Leadership and ESG Performance: A Meta-Analysis

Motivation Public companies are facing growing pressure from investors and regulators alike. By 2026, certain large companies that do business in the EU will have to adopt a fundamental pillar of the broad "EU Green Deal", EU's Corporate Sustainability Reporting Directive (CSRD), which requires mandatory Environmental, Social and Governance (ESG) reporting - shedding more light on the importance of the firm's ESG performance itself. In addition, those publicly listed of them will simultaneously have to take measures to increase women's presence at their helm, according to the so-called "Women on Boards" Directive. But what if increasing the female representation on boards could kill two birds with one stone.

According to many contemporary articles, females are expected to be more ESG oriented. This may arise from their natural ability to exhibit higher rates of empathy-related traits (Christov-Moore et al., 2014). When it comes to professional performance, whether the nurturing social cognition dominates or not, female leaders are shown to be more socially responsible (Borghesi et al., 2014; Byron and Post, 2016; Eagly, 2009; Hofstede, 1998). At the same time, gender-diversified boards are more likely to propose alternative solutions, manage firm's social and environmental challenges (Shaukat et al., 2016), being committed to ethics and consider the interests of a wide range of stakeholders and communities (Kennedy and Kray, 2014).

Recent organizational research already aims to assess the intersection of female leadership and environmental, social and governance performance. Yet, the empirical results are mixed. Even though the positive estimated relationship is presented in most of the studies examined (Alkhawaja et al., 2023; Birindelli et al., 2018; Khatri, 2023; Nadeem et al., 2017; Velte, 2016), some studies still suggest a negative or no relationship at all (Cucari et al., 2018; Husted and Sousa-Filho, 2019; Manita et al., 2018). One of the reasons that findings across studies may vary is that studies tend to focus on firms

from a single country. In addition, as found by Aabo and Giorici (2023), the importance of females in the executive board for the firm's ESG profile depends crucially on the data used for the analysis.

Thus, we see a lot of potential in performing the meta-analysis on the topic of female leadership and ESG performance as it allows us to determine the true variability in the specified relationship by taking advantage of the causal effects that have been presented in the recent literature and demonstrate the direction and extend of the potential bias. As a result, large companies that concurrently fall under the Corporate Sustainability Reporting and "Women on Boards" directives may benefit from this study when making decision in adaptation of the new regulations and compliance with EU standards.

Hypotheses

1. Hypothesis 1: The literature on female leadership and ESG performance is affected by publication bias.

2. Hypothesis 2: Publication bias exaggerates the mean value of the estimated effect of female leadership on ESG performance.

3. Hypothesis 3: The estimated effect of female leadership on ESG performance is driven by geographical location.

Methodology Following the methodology of a typical meta-analysis, set of primary studies will be collected in the first place. This will be done by defining a search query and using the Google Scholar database to search for the full-text studies. The focus will be given on the studies reporting standard errors, number of observations, standard deviation, confidence intervals and other effects relevant to the subsequent analysis of the heterogeneity. Once the dataset is acquired, the publication bias will be examined using various approaches. At the beginning, it is usually beneficial to start with graphical representation of the publication bias using the funnel plot as described by Egger et al. (1997). In addition to graphical test, which may lack statistical validity, several other tests will follow, including the Funnel Asymmetry Test (FAT) that employs different estimators and weighting matrices and was firstly proposed by Stanley (2005). Acknowledging the potential non-linearity in the relationship between the effect and the standard error, various non-linear methods will be applied too, including the Top10 approach (Stanley et al., 2010), the weighted average of adequately powered tests proposed by Ioannidis et al. (2017), Furukawa (2019) stem-based method, Andrews and Kasy (2019) selection model based on Hedges (1992), and the endogenous kink method by Bom and Rachinger (2019). Additionally, the p-uniform* test will be applied (van Aert and van Assen, 2018), as it utilizes p-values to identify publication bias without assuming a specific relationship between the estimate and its standard error.

In the subsequent part of the thesis, the focus will shift to examining heterogeneity. To accomplish this, Bayesian Model Averaging (BMA) will be employed (Steel, 2020), incorporating the dilution prior suggested by George (2010) to address multicollinearity in the data. As an additional robustness check, Frequentist model averaging will be used (Amini and Parmeter, 2012).

Expected contribution Up till now, no meta-analysis has been performed on the topics of female leadership and ESG performance even though the recent research provides numerous publications on the subject of the matter. As suggested, ESG and gender diversity topics are still sustainable target of not only investors' but also regulators' attention. Providing a comprehensive summary and unbiased estimate on the effect of female in the boards on the company's ESG performance may provide insights into how organizations can leverage diversity for strategic advantage in the aftermath of the recent directives of the European Parliament and the Council of the European Union.

Outline

1. Introduction: Main motivation and expected contribution will be presented.
2. Literature Review: Available literature will be summarized - existing surveys and their results on our topic of interest will be reviewed.
3. Data Selection: Selection criteria and a simple summary statistic of the final dataset will be provided and discussed.
4. Estimation of the Causal Relationship: The estimation methods employed and the proxies for dependent variable (ESG performance) and variable of interest (female leadership) among publications in our data will be summarized.
5. Methodology:
   (a) Publication bias: Methods for publication bias testing will be introduced and employed.
   (b) Heterogeneity: Methods for testing and identification of major sources of heterogeneity will be introduced and employed.
6. The best-practice estimate
7. Conclusion

Core bibliography

Aabo, T. and Giorici, I. (2023). Do female CEOs matter for ESG scores? *Global Finance Journal*, 56. Publisher: Elsevier B.V.

Alkhawaja, A., Hu, F., Johl, S., and Nadarajah, S. (2023). Board gender diversity, quotas, and ESG disclosure: Global evidence. *International Review of Financial Analysis*, 90. Publisher: Elsevier Inc.

Amini, S. and Parmeter, C. (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, 27.

Andrews, I. and Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–94.

Birindelli, G., Dell'Atti, S., Iannuzzi, A., and Savioli, M. (2018). Composition and activity of the board of directors: Impact on ESG performance in the banking system. *Sustainability (Switzerland)*, 10(12). Publisher: MDPI.

Bom, P. and Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, 10(4):497–514. Publisher: John Wiley and Sons Ltd.

Borghesi, R., Houston, J., and Naranjo, A. (2014). Corporate socially responsible investments: CEO altruism, reputation, and shareholder interests. *Journal of Corporate Finance*, 26:164–181. Publisher: Elsevier.

Byron, K. and Post, C. (2016). Women on Boards of Directors and Corporate Social Performance: A Meta-Analysis. *Corporate Governance: An International Review*, 24(4):428–442. Publisher: Blackwell Publishing Ltd.

Christov-Moore, L., Simpson, E., CoudĂ©, G., Grigaityte, K., Iacoboni, M., and Ferrari, P. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience and Biobehavioral Reviews*, 46(P4):604–627. Publisher: Elsevier Ltd.

Cucari, N., Esposito De Falco, S., and Orlando, B. (2018). Diversity of Board of Directors and Environmental Social Governance: Evidence from Italian Listed Companies. *Corporate Social Responsibility and Environmental Management*, 25(3):250–266. Publisher: John Wiley and Sons Ltd.

Eagly, A. (2009). The His and Hers of Prosocial Behavior: An Examination of the Social Psychology of Gender. *American Psychologist*, 64(8):644–658.

Egger, M., Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634. Publisher: BMJ Publishing Group.

Furukawa, C. (2019). Publication Bias under Aggregation Frictions: Theory, Evidence, and a New Correction Method. *SSRN Electronic Journal*.

George, E. (2010). Dilution Priors: Compensating for Model Space Redundancy. *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, 6.

Hedges, L. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255.

Hofstede, G. (1998). Attitudes, values and organizational culture: Disentangling the concepts. *Organization Studies*, 19(3):477–492. Publisher: Walter de Gruyter and Co.

Husted, B. and Sousa-Filho, J. (2019). Board structure and environmental, social, and governance disclosure in Latin America. *Journal of Business Research*, 102:220–227. Publisher: Elsevier Inc.

Ioannidis, J., Stanley, T., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605):F236–F265. Publisher: Blackwell Publishing Ltd.

Kennedy, J. and Kray, L. (2014). Who Is Willing to Sacrifice Ethical Values for Money and Social Status?: Gender Differences in Reactions to Ethical Compromises. *Social Psychological and Personality Science*, 5(1):52–59.

Khatri, I. (2023). Board gender diversity and sustainability performance: Nordic evidence. *Corporate Social Responsibility and Environmental Management*, 30(3):1495–1507. Publisher: John Wiley and Sons Ltd.

Manita, R., Bruna, M., Dang, R., and Houanti, L. (2018). Board gender diversity and ESG disclosure: evidence from the USA. *Journal of Applied Accounting Research*, 19(2):206–224. Publisher: Emerald Group Publishing Ltd.

Nadeem, M., Zaman, R., and Saleem, I. (2017). Boardroom gender diversity and corporate sustainability practices: Evidence from Australian Securities Exchange listed firms. *Journal of Cleaner Production*, 149:874–885. Publisher: Elsevier Ltd.

Shaukat, A., Qiu, Y., and Trojanowski, G. (2016). Board Attributes, Corporate Social Responsibility Strategy, and Corporate Environmental and Social Performance. *Journal of Business Ethics*, 135(3):569–585. Publisher: Springer Netherlands.

Stanley, T. (2005). Beyond publication bias. *Journal of Economic Surveys*, 19(3):309–345. Publisher: Blackwell Publishing Ltd.

Stanley, T., Jarrell, S., and Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *American Statistician*, 64(1):70–77.

Steel, M. (2020). Model Averaging and Its Use in Economics. *Journal of Economic Literature*, 58(3):644–719. Publisher: American Economic Association.

van Aert, R. C. M. and van Assen, M. A. L. M. (2018). Correcting for publication bias in a meta-analysis with the P-uniform* method. Publication Title: BITSS.

Velte, P. (2016). Women on management board and ESG performance. *Journal of Global Responsibility*, 7(1):98–109. Publisher: Emerald Group Holdings Ltd.

Author

Supervisor

Chapter 1

Introduction

In recent years, research on the relationship between female leadership and Environmental, Social and Governance (ESG) performance has experienced a significant surge. The growing availability of comprehensive ESG data, coupled with an increasing representation of women in leadership positions has sparked considerable interest among scholars. A seminal study in this area is that of Manita et al. (2018), one of the most frequently cited papers, which models the relationship between female leadership and ESG performance. Focusing on US firms between 2010 and 2015, authors report an effect not significantly different from zero. Two years later, in 2020, Shakil (2021) revisited the relationship between the two variables on a sample of US banks using a more recent dataset spanning from 2013 to 2017. Contrary, a significant positive effect is reported with an average magnitude of approximately 0.5 points, quickly garnering citations of Manita et al. (2018). This trend, however, extends beyond US banks. Similar publications emerge in sectors such as energy (Shahbaz et al., 2020), healthcare (Uyar et al., 2021) or even extractive industries (Wang et al., 2022), as well as in regions like Latin America (Husted and de Sousa-Filho, 2019) or the Middle East (Issa et al., 2022).

Based on 547 estimates from 108 studies, a 1% increase in board gender diversity corresponds to an average increase of 0.24 points in ESG scores. However, the estimated effects of gender diversity show considerable variation, ranging from -0.33 to 5.9 points. This raises important questions: What accounts for this wide dispersion in the reported effects? Is it driven by factors such as year of data, sample size, methodology or geographical location? And importantly, what is then the true effect of board gender diversity on ESG performance? These questions are not just of academic interest. The answers have significant implications for both policy-making and corporate governance. In this context, after over a decade

of research, there is now sufficient evidence to formally examine these questions through meta-analysis, a quantitative method of research synthesis that enables a comprehensive and systematic evaluation of the available studies.

Meta-analysis is a valuable method for combining results from multiple studies, offering more robust and reliable conclusions. To the best of our knowledge, no existing meta-analysis has directly examined the effect of female directors on ESG performance. The two studies that come closest are those by Wu et al. (2022) and AlJanadi (2025) exploring gender diversity effects on Corporate Social Responsibility disclosures. In case of Wu et al. (2022), analysis of 44 papers suggests that firms with greater female board representation tend to have better CSR outcomes, but not necessarily higher ESG scores. While both studies offer valuable insights, they do not thoroughly address publication bias or investigate the drivers of heterogeneity across estimates. In contrast, our meta-analysis applies the most recent meta-analytical techniques such as FAT-PET, weighted average of adequately powered due to Ioannidis et al. (2017), Selection model proposed by Andrews and Kasy (2019), Stem-based technique suggested by Furukawa (2019), endogenous kink model of Bom and Rachinger (2019) and p-uniform* technique due to van Aert and van Assen (2021). In addition, to identify drivers of the heterogeneity among the estimated effect, Bayesian model averaging and frequentist check are applied. Our key finding is that the relationship between board gender diversity and ESG ratings is not causal. The average effect of 0.24 points reported in the literature is inflated due to publication bias. After applying conservative corrections for publication bias, the mean effect is reduced to approximately 0.1 points. Consequently, application of more aggressive methods results in an even greater reduction. Furthermore, when controlling for the presence of the CSR committee in the primary study or using panel methods, the remaining effect of female leadership, after correcting for publication bias, decreases to nearly zero.

The structure of this thesis is as follows. Chapter 2 sets the stage by providing the conceptual framework necessary to explore ESG and female leadership. In Chapter 3, we present a summary of relevant studies that examine the relationship under consideration, their estimation techniques and measurement of the variables of interest. Chapter 4 outlines the data collection procedure, discusses the data adjustment process and presents the descriptive statistics of our sample. Chapter 5 is dedicated to assessing publication bias. In Chapter 6, we introduce Bayesian model averaging, examine the factors that contribute to the heterogeneity across existing estimates and present the best practice estimates. Finally, Chapter 7 concludes the thesis. The dataset and the code used are available upon request.

Chapter 2

Conceptual Framework

2.1 Female Leadership Trends

In 2020, governments, non-governmental organizations, associations, companies and investors were brought to action by an intense wake-up call. The results of World Economic Forum (2020) suggested that it will take roughly 100 years for women to achieve global equality. As a result, considerably more attention has been placed at uniform contribution of women and men, equal opportunities at the workplace and, to a lesser extent, hiring rates of women into senior leadership roles. Now, four years later, the gap is almost bridged among entry-level positions, however, the ascent to the highest executive roles remains tough (World Economic Forum, 2024). Only about 6% of chief financial officers are women (Diligent Institute, 2022; Equileap, 2024; European Women on Boards, 2024; Csonka and Milhomem, 2023), and the percentage of women chairing boards is reported to fluctuate somewhere between 8 to 10 percent globally, nearly three times lower than the proportion of board seats occupied by women (Deloitte Insights, 2023; Diligent Institute, 2022; Csonka and Milhomem, 2023).

A wide range of different sources unanimously claim a global positive course in the direction of higher gender parity among leadership positions (Diligent Institute, 2022; Equileap, 2024; European Women on Boards, 2024; Csonka and Milhomem, 2023), however, the distribution of the progress is uneven across various countries and industries. Information technology, infrastructure or energy industries constantly report the lowest female participation compared to sectors such as healthcare and education, where women are relatively better represented among leadership roles (Deloitte Insights, 2023; Diligent Institute, 2022; European Institure for Gender Equality, 2023; European Women on Boards, 2024; Csonka

and Milhomem, 2023; World Economic Forum, 2024). Although th education sector seems to offer relatively easy access to top management positions, the most significant drop-to-top yield real-estate and financial industries (World Economic Forum, 2024).

The variation also increases with different countries. On average, there are more women directors in developed countries than in emerging ones (European Institure for Gender Equality, 2023; Equileap, 2024; Csonka and Milhomem, 2023). The regulatory environment and the pressure from investors could be assumed as factors that represent significant differences. However, it is also the political system and cultural settings of the country that forms women's role within the society and herewith the occupation of leadership positions (Carrasco et al., 2015). For example, South Africa, despite being a developing country without mandatory gender quotas, exhibits a relatively high share of women on boards, not far behind European countries such as France, Norway, and Italy, countries which adopted gender quotas years ago and now lead globally, with female board representation over 40% (European Women on Boards, 2024; Csonka and Milhomem, 2023). In contrast, countries such as Qatar and Saudi Arabia have not yet adopted any measures to increase female board representation, and are repeatedly on the tail end in female leadership (Csonka and Milhomem, 2023).

While some regions struggle to initiate change, others pursue alternative strategies to enhance board diversity. In the United States, a markedly different landscape emerges. Although there is no federal specific mandate for gender diversity on corporate boards, several states have enacted laws to promote female representation. These efforts are complemented by pressure of institutional investors and advocacy by non-governmental organizations. Major asset management firms like BlackRock, Vanguard Asset Management and State Street Global Advisors have policies to vote against nominating committee members of boards that lack diversity. Moreover, in 2021 the Securities and Exchange Commission (SEC) approved NASDAQ's rule requiring companies listed on its US exchange to have at least one woman on board. The New York Stock Exchange (NYSE), in contrast, promotes board diversity through its Board Advisory Council, which leverages CEOs' networks to connect diverse candidates with companies. This network-driven approach focuses on expanding access to personal networks for diverse candidates rather than enforcing direct diversity requirements.

2.2 Topics in ESG

In recent years, tension over sustainability became almost tangible. Investors call for more sustainable investments, while consumers and employees look for brands that align with their values. Companies appeared to be under strong pressure to operate responsibly, invest sustainably and provide meaningful benefits to a broad range of stakeholders.

However, the concept of sustainable investment is not a novelty of business practices. In reality, this notion appeared in the 1990s when investors began to establish Corporate Social Responsibility (CSR) committees to identify firms with common values (Puaschunder, 2019). In a way, CSR may be seen as a precursor of Environmental, Social and Governance (ESG). The transition from CSR to ESG was largely driven by the demand for standardized, comparable data that could help investors assess sustainability performance in financial terms. While CSR is mainly described as companies' efforts and voluntary actions to contribute positively to society, ESG offers more concrete and quantifiable criteria. The first of the three pillars that ESG is based on, the environmental one, encompasses indicators such as efficiency in energy use, reduction of waste, greenhouse emissions or carbon footprint. Secondly, the social pillar provides information on living wages, equal employment opportunities, employee benefits, workplace health and safety, community engagement or responsible supply chain partnerships. Third, the governance pillar underlines corporate governance in risk management, compliance, ethical business practices, transparency and integrity.

In a nutshell, ESG represents environmental, social and governance values. Thus, ESG performance should provide quantification of the degree to which a company adheres to these values. Although the performance of most of the ESG indicators can be numerically represented (quantitative ESG indicators), there are still some actions that are more subjective and thus hard to measure, such as the strategy for the loss of biodiversity due to climate change or the impact on local communities. These factors are also called qualitative metrics and often require further interpretation. In addition, there is no one size fits all. That is the reason why relevant issues for ESG differ among industries. However, these would not be the most critical issues. The most problematic appears to be the disclosure of such values. Firms must first become transparent about their path to sustainability in order to quantify this journey. In other words, companies must bridge the informational asymmetry gap for their stakeholders by reporting ESG strategies and activities. This is usually done by publishing the so-called

Environmental, Social and Governance Disclosure (ESGD) that summarizes non-financial strategies and actions toward sustainability. ESGD is a good indicator that the company values not only economic profits but also the environment, society, consumers, and employees (Abdul Rahman and Alsayegh, 2021).

Recently, strong pressure from investors has further amplified the importance of ESGD. As sustainable investment has transitioned from a niche market to the mainstream, many managers have recognized a simple financial reality: the better a firm's ESG reputation, the lower its funding costs (Wong and Zhang, 2022). Consequently, sustainability disclosure is no longer merely about ethics, but has become a strategic tool. Li and Wang (2022) even notes that financial incentives can trigger synergistic effects, encouraging firms to make more substantial commitments to corporate social responsibility. However, while reporting may reflect genuine progress, it is often strategically leveraged to secure financing and build a positive public image (Berrone et al., 2017; Delmas and Burbano, 2011).

This strategic use of disclosure has intensified the role of ESG ratings in defining what constitutes "sustainable enough". ESG rating agencies are generally transparent about their methodologies, enabling managers to tailor their disclosures according to established frameworks such as Global Reporting Initiative (GRI) and Sustainability Accounting Standards Board (SASB). Yet, such practices frequently give rise to greenwashing, the act of portraying corporate sustainability efforts in a misleadingly favorable light (Zhang, 2022, 2023). Growing concerns about greenwashing have led to increased regulatory scrutiny, with initiatives such as the EU's Green Claims Directive aiming to enhance transparency and combat misleading sustainability claims.

2.3 ESG Ratings

Speaking of ratings, there is currently a large number of ESG data providers. At the same time, with the same mission - to enable comparable evaluations of ESG performance data. Nevertheless, ESG ratings date back to 1983 when Eiris, the first ESG rating agency, was established in France. Seven years later, the USA came with Kinder, Lydenberg, and Domini (KLD). As the topic of sustainable investing grew, the market for ESG ratings did too. Furthermore, several early ESG rating firms were acquired or merged with other financial data companies: Morgan Stanley Capital International (MSCI) purchased KLD in 2010, Eiris merged with Vigeo in 2015, Morningstar obtained a 40% stake in Sustainalytics in 2017, then Moody's acquired Vigeo-Eiris in 2019, and S&P Global took over RobecoSAM.

Today, the most prominent ESG data providers are MSCI, S&P Global, Sustainalytics, Institutional Shareholder Services (ISS), Bloomberg, London Stock Exchange Group (LSEG) (formerly Refinitiv or ASSET4) and Moody's. ESG ratings provide a service to investors by collecting and aggregating information from a wide range of sources and reporting standards. Adhering to their own methodologies, frameworks and computation processes, while strongly relying on the ESG disclosures, each rating agency computes its own ESG score. While MSCI and ISS scores ESG on a scale 1 (or AAA, the highest ESG Rating) to 10 (or CCC, the lowest ESG score), Bloomberg, LSEG and S&P Global utilize a 0.01-100 point scale, the highest score the better ESG performance. Sustainalytics uses the opposite logic, with their ratings being closer to 100, the higher the probability of ESG risk.

Reasonably, resemblance is not all that matters. In particular, many scholars have shown that scores disagree remarkably and the correlations between ESG ratings of different rating agencies remain low (Berg et al., 2022; Capizzi et al., 2021; Chatterji et al., 2016; Dorfleitner et al., 2015). Chatterji et al. (2016) names two main reasons for the divergence of ESG scores - what is measured and whether it is measured consistently. Berg et al. (2022) elaborates on the sources of divergence and suggests that the ESG ratings differ due to different sets of attributes used for the calculation, different measurement and different relative weights given to the same attribute.

The divergence of ESG ratings has various implications. It affects not only the risk-return trade-off (Avramov et al., 2022), income fluctuations, and external financing (Christensen et al., 2021), but also the validity of empirical studies based on a single ESG rating (Dorfleitner et al., 2015). Capizzi et al. (2021) points out that certain empirical results based on one type of ESG rating may not be replicable when using ESG ratings from a different provider.

As divergence in ESG ratings has been a persistent issue, raising concerns not only among investors but also among policymakers and researchers (Chatterji et al., 2016; Gibson Brandon et al., 2021), authorities have intervened. The EU's Corporate Sustainability Reporting Directive (CSRD), which replaces the Non-Financial Reporting Directive (NFRD), introduces stricter requirements for sustainability disclosures and explicitly defines a standardized reporting framework. This initiative is expected to make companies' ESG disclosures more reliable, transparent, and standardized, reducing ambiguity in ESG score assessments.

2.4 Exploring ESG Through Leadership

Once data sees the light of day, it would be a pity not to use it. At least that is what researchers would agree on. Not surprisingly, academic journals have been swarmed with publications using ESG scores from various sources. Scholars do not only assess whether ESG resonates with investors, consumers or employees, but reasonably also try to provide an impression of what could stimulate the ESG performance itself. Besides country-level indicators such as reporting regulations (Lokuwaduge and Heenetigala, 2017), stock exchange membership (Weber, 2014), corruption (Baldini et al., 2018; Hoang, 2022; Khalid et al., 2022), democracy and political stability (Mooneeapen et al., 2022), there are many company-related aspects that play a significant role. Profitability, size, leverage, and tangibility (Özcan, 2019; Abdul Rahman and Alsayegh, 2021), but also firm ownership (Ellili, 2023; Weber, 2014). However, the attributes of top management belong to one of the determinants of ESG. CEO's narcissism, age, compensation, tenure or duality, as well as board characteristics such as size, independence or cultural and gender diversity are being linked to firm's ESG performance (Arayssi et al., 2020; Chebbi and Ammer, 2022; Dabbebi et al., 2022; Ellili, 2023; Gurol and Lagasio, 2022; Husted and de Sousa-Filho, 2019; Lavin and Montecinos-Pearce, 2021; McBrayer, 2018).

This strongly relates to the upper echelons theory developed by Hambrick and Mason (1984) who argues that the best way to understand how a firm performs is to discover the fundamental biases and tendencies of the top management. In other words, executive attributes can be used to predict organizational outcomes (Shahab et al., 2018).

2.5 Why Women Directors Matter for ESG

The most pronounced determinant of business destiny remains the effectiveness of its board (Drucker, 2012). The main role of the board of directors is to protect shareholders' interest. That is why board effectiveness is mainly determined by the monitoring and strategic advisory skills of the directors (Adams et al., 2010; Forbes and Milliken, 1999; Minichilli et al., 2012). Strategic control tasks, which can also involve the supervision of managerial decisions concerning organizational direction in relation to safety, health, and environment, are better managed by diverse boards (De Andres et al., 2005; Nielsen and Huse, 2010). Women may contribute unique and valuable perspectives on human and relational capital that is

an important building block to master the monitoring ability of the board (Adams and Ferreira, 2009; Hillman et al., 2007). Female directors behave more responsibly when it comes to attending board meetings and promoting them among their male peers (Adams and Ferreira, 2009), resulting in improved board dynamics (Konrad et al., 2008).

In addition, theories of gender socialization and ethics suggest that women's nurturing and empathetic potential is deeply embedded in their nature (Gilligan, 1977). When it comes to the expression of these characteristics in their professional life, the literature indicates that women are more inclined to endorse social values that benefit others and are less likely to support unethical decisions (Borkowski and Ugras, 1998; Boulouta, 2013; Franke et al., 1997; Schwartz and Rubel, 2005; Williams, 2003). Empowering women in leadership positions is associated with more policy making that addresses the concerns of families, women, and ethnic and racial minorities, as well as improvements in standards of living, social gender equality, and national wealth (Beaman et al., 2009; Cohen and Huffman, 2007; Eagly et al., 2014). Research has shown that women employees under female leadership address less gender inequality (Konrad et al., 2010), lower pay disparity (Shin, 2012) and more organizational and family support (Foley et al., 2006).

As women often show a greater focus on society's broader challenges compared to men (Atif et al., 2021), their presence in the boardroom helps to manage the relationships of their key stakeholders (Brammer et al., 2015). Women can understand the needs and support the expectations of various stakeholder groups (Amorelli and García-Sánchez, 2021; Ibrahim and Hanefah, 2016), as well as respond effectively to their interests (Harjoto et al., 2015), even in terms of environmental and social issues (Bear et al., 2010; Ben-Amar et al., 2017; Mackenzie, 2007; Rao and Tilt, 2016).

Whether it is caring intuition or strategic endowment that dominates, there exist many publications suggesting female leaders are more prone to advocate for climate action and sustainability initiatives. Moreover, they are twice as likely to exhibit sustainable experience compared to their male counterparts (Diligent Institute, 2022). Women-led companies prioritize proper waste management, make efforts to reduce the misuse of emissions and resources, invest more in renewable energy, allocate less to polluting industries, and consequently perform better on ESG indicators (European Investment Bank, 2022; Pavlova and Gvetadze, 2023; Pierli et al., 2022).

Chapter 3

Notes on Estimation Techniques

In terms of research on the effect of female leadership, the results show a great deal of variety. Speaking of US companies, Aabo and Giorici (2023) demonstrate that whether female CEO makes a difference to firm's ESG profile strongly depends on the data used. In their analysis of the same sample of firms, employing data provided by Refinitiv resulted in no association between CEO's gender and sustainability performance, while based on Bloomberg data the relationship is significant and positive. On the other hand, Chinese companies show no effect of female CEO on ESG scores even when using Bloomberg's scores (Li et al., 2023; Liu et al., 2024). Quite weak and rather insignificant relationship is found among UK companies too, when the Bloomberg's ESG data is used (Harjoto and Wang, 2020). On the other side, women's representation on audit yields positive impact on CSR, as shown on the LSEG sample of US and French companies (Boukattaya et al., 2024; Yorke et al., 2023). However, if it is a woman who chairs the board among Italian listed companies, the performance of ESG increases only until the critical mass of female directors in the boardroom is reached (Słomka-Gołębiowska et al., 2023).

Even more inconclusive remains the research on the impact of female board directors. Similarly to CEO's gender, Martiny et al. (2024) claim that impact of board diversity depends on the ESG data provider used. However, there might be various other factors that imply the relationship. Recent global evidence provided by Alkhawaja et al. (2023) suggests that companies with gender-disparous boards disclose more ESG-related topics to their stakeholders in 48 countries. In addition, the authors suggest that following the implementation of gender quota at the national level, the ESG scores among firms increase significantly. The estimation of the positive effect remains consistent among samples of various countries

and industries. Studies focusing on listed firms in the Gulf Cooperation Council (Arayssi et al., 2020), France (Nekhili et al., 2021), Malaysia (Wasiuzzaman and Wan Mohammad, 2020) or Germany and Austria (Velte, 2016), as well as samples from energy firms (Shahbaz et al., 2020), banks (Birindelli et al., 2018) or companies engaged in tourism and hospitality (Uyar et al., 2020) provide additional evidence of the diversity of the board gender in relation to ESG. In contrast, Manita et al. (2018) finds an insignificant relation between female board participation and ESGD in 379 US companies over the period 2010 to 2015. Using samples from Italian and Latin American companies over almost the same time period (2011-2014), Cucari et al. (2018) and Husted and de Sousa-Filho (2019) report a significantly negative impact of women on boards on ESG performance, which is argued to be the result of a small proportion of women on corporate boards. In summary, while positive estimates of female leadership dominate, the true effect of women on boards remains questioned because the regulatory environments, markets and social norms across countries are different.

3.1 Estimation Methods

Furthermore, part of the heterogeneity among the estimated effects may be driven by different estimation methods that scholars employ. Usually, authors assess the effect of female directors on ESG performance using various alternatives of the following model:

$$ESG_{it} = \beta_0 + \beta_1 GD_{it} + X_{it} + \epsilon_{it} \tag{3.1}$$

where ESG_{it} represents the ESG performance of firm i in time t; GD_{it} is a proxy of gender diversity; X_{it} corresponds to a vector of control variables utilized in the study and ϵ_{it} refers to the error term.

The existing literature struggles with various sources of endogeneity when estimating Equation 3.1. Most publications utilize ESG scores that are calculated by evaluating various environmental, social and governance factors. Firms understand that they can intrinsically improve their performance by ma-king simple adjustments to the issues taken into account when constructing ESG ratings. In this case, companies can hire more female directors to improve their governance score, which is, in turn, one of the components in the calculation of the ESG performance. Speaking of reverse causality, it is mainly present in cross-sectional data where the two variables, female leadership and ESG performance, are observed

simultaneously in time t. Consequently, the use of panel data is helpful, as the female leadership observed at time t can be used as a predictor of the outcomes of ESG measured at time $t + 1$.

In addition, publications may suffer from omitted variable bias or unobserved firm characteristics related to both ESG and female leadership at the same time (Adams and Ferreira, 2009). For example, companies with progressive corporate culture might invest more in their ESG practices while simultaneously encouraging gender diversity in leadership. Furthermore, estimation of the causal relationship faces selection bias when the assignment of female directors to boards is no longer random but is driven by external factors such as quotas or female directors' preferences, such as self-selection into firms that are more socially responsible and align with values such as diversity and inclusion (Atif et al., 2021).

Considering a wide range of endogeneity that leads to a spurious relationship between female leadership and ESG performance, simple ordinary least squares (OLS) estimator might not always produce a reliable and consistent estimate. Nevertheless, OLS method is still applied in few studies trying to model the relationship between the two variables. Studies, however, differ in their approaches. While some publications utilize pooled OLS with longitudinal data and disregard the panel structure of the data, others simply include dummies for year, country and sector to account for individual and time-specific effects (Alkurdi et al., 2023). In addition, some of them simultaneously try to address reverse causation by applying lagged values of independent variables (Alkhawaja et al., 2023). Despite this, there may be some unobserved firm-level characteristics that do not change over time but impact both firm ESG and diversity in the boardroom, such as corporate culture. Similarly, the usage of pooled OLS may be unfeasible when dealing with large panel data. Inclusion of many dummy variables for fixed effects may result in multicollinearity, which inflates standard errors and reduces the precision of estimates.

Therefore, the vast majority of studies rely on fixed effects (FE) estimators. Due to the fact that the FE method relaxes the strict exogeneity assumption and allows for correlation between the fixed effects and the independent variables, FE is considered to be a more robust estimator in the presence of such correlations. Moreover, it effectively controls for time-invariant and firm-specific unobserved heterogeneity. Therefore, FE estimates are expected to be more reliable than those obtained through OLS. However, the fixed effects model may still fail to address other possible sources of endogeneity. Endogenous relationship between female board representation and ESG rating can be overcome by employing lagged

values, however, the dynamic endogeneity caused by the inclusion of past values of the firms ESG as an explanatory variable may not be fully addressed in the FE model.

Few authors prefer the random effects (RE) model over the fixed effect model, but it is rather exceptional. For example, Adamu et al. (2024) analyzed the relationship between female directors and ESG score among Nigerian banks. With 88 firm-year observations over eight years, their dataset encompasses relatively few observations per firm. In such a case, it is reasonable to favor the random effects model or fixed effects model over pooled OLS. The RE model can use both the within-firm and between-firm variation, improving efficiency in small datasets. Thus, in case strict exogeneity assumption is met, random effects model leads to smaller standard errors and more precise estimates.

Due to reverse causality, the smaller part of the explored studies employ two-stage least squares (2SLS). The effectivity of the 2SLS approach lies in its ability to extract an exogenous component of the endogenous variable in the first stage, which is subsequently used to estimate the effect of female leadership on ESG performance. The issue of correlation between the endogenous variable and the error term can be addressed by selecting a suitable instrumental variable, often referred to as an instrument, which must fulfill two key conditions. It must be relevant (meaning it is correlated with the endogenous variable) and exogenous (meaning it is uncorrelated with the error term of the structural equation). If either of these conditions is not met, the resulting estimates may be inconsistent (Antonakis et al., 2010). Most often, primary studies instrumentalize the representation of women on boards by the size of the board itself (Dang et al., 2023a; Trireksani et al., 2024). Other examples of instruments used are board independence (Khemakhem et al., 2023), industry's average women proportion excluding the firm itself and women proportion on boards in the initial period (Khatri, 2023) or firm size (Nuhu and Alam, 2024). In contrast, an unsuccessful search for a valid external instrumental variable in the study of Issa et al. (2022) resulted in the decision not to employ the 2SLS technique.

Consequently, the authors prefer generalized methods of moments (GMM). Considering the wide range of bias that can be encountered when modeling the effect of board gender diversity on corporate sustainability, GMM shows to be a safe place. Classic GMM provides a general framework for models with endogenous variables, heteroskedasticity and autocorrelation in errors at once. This method takes advantage of lagged lagged independent variables from which instruments are produced internally. Subsequently, few extensions of the GMM

technique are employed also within the primary studies of interest. A two-step GMM is a method commonly used among scholars (Dang et al., 2023b; Gerged et al., 2023; Nekhili et al., 2021; Issa et al., 2022). In contrast to classic GMM, it iteratively refines the weighting matrix and produces more efficient estimates. Furthermore, few studies employ a dynamic version of GMM (Benaguid et al., 2023) and a system GMM to reduce the issue of endogeneity, measurement error, bias of omitted variables, and firm-specific heterogeneity (Dang et al., 2021; Shakil, 2021).

3.2 Measurement of Female Leadership

In the literature, female leadership and gender diversity in senior management are commonly captured through different proxy, such as the share of women on boards, a dummy variable indicating a female CEO, indicators for reaching a specific number of women directors, the absolute count of women in executive roles or various diversity indices.

The most widely used measure of gender diversity is the proportion of women in the boardroom (Alkhawaja et al., 2023). This approach is widely accepted, offering a straightforward and quantifiable measure of female representation, typically expressed as a fraction, ratio or percentage. It facilitates straightforward comparisons between companies and industries while allowing to track changes over time.

Another commonly used variable is the dummy variable method that assigns a value of 1 in the event that at least one woman in the boardroom is present and 0 otherwise (Aladwey and Alsudays, 2023; Chebbi and Ammer, 2022). However, to gain a more nuanced understanding of the impact of female leadership on ESG performance, researchers often apply the critical mass theory (De Masi et al., 2021). The theory proposes applying a stricter criterion for female representation by setting the dummy variable to 1 only if there are at least three women in leadership positions, and 0 otherwise (Cambrea et al., 2023).

Furthermore, the Blau (1977) index and the Shannon (1948) index are often utilized as robust methods to measure gender diversity (Abdullah et al., 2024; Gangi et al., 2021; Fayyaz et al., 2023; Romano et al., 2020). The Blau index is calculated using the following formula:

$$1 - \sum_{i=1}^{k} p_i^2 \tag{3.2}$$

where k denotes the total number of individuals on the board or top management team, and p_i reflects the proportion of each gender within the group. The Shannon index is then calculated as follows:

$$-\sum_{i=1}^{k} p_i ln(p_i) \tag{3.3}$$

where k and p_i are defined as in the Blau index. Due to the Shannon Index's reliance on logarithms and the Blau Index's quadratic function, they are inherently more sensitive to smaller values, making them potentially better suited for capturing critical mass effects. However, a limitation of the logarithm-based metric is that it becomes undefined when p_i equals 0. The indices range from 0 (when all members belong to a single gender) to 0.5 or 0.693 (when equality is reached) for the Blau and Shannon indices, respectively.

3.3 Measurement of ESG Performance

Similarly to female leadership, the dependent variable can take various forms. Comparing ESG performance between companies or tracking it over time within a single company is challenging. Similarly to investors, the authors of the studies often grapple with the challenge of its measurement.

Content analysis, also known as a textual analysis of corporate sustainability reporting (Ferjančič et al., 2024), seems to be a popular choice among many scholars who do not want to rely on ESG scores assessed by third parties. As the name may already suggest, this method for the computation of own ESG scores usually involves analyzing the content of company's financial or sustainability disclosures and verifying the inclusion of specific measures. Some studies do not necessarily rely only on the information provided in the disclosures, however, they also actively include information provided by independent publicly available sources (Farooque et al., 2022). Subsequently, a point is assigned for each case where a company or other stakeholder reports on a specific ESG indicator. The final ESG score is then computed as a fraction of ESG indicators company complies with over all ESG indicators considered in the analysis. For example, Bel-Oms and Segarra-Moliner (2022) looked into 140 environmental, social and economic factors when constructing their ESG index.

Due to the burdensomeness this method brings, it is reasonable that employment of artificial intelligence and natural language processing would be of help. Sepulveda-Nuñez et al. (2024) and Sutiono (2020) utilized data generated by Tru-

Value Labs that is based on the continuous collection, interpretation and evaluation of ESG indicators across various sources, promising transparent and nearly real-time ESG performance data.

Lastly, and most commonly, vast majority of studies rely on ESG scores computed by various ESG rating agencies, mainly by Bloomberg and LSEG (former Refinitiv or ASSET4), but also MSCI (Liu et al., 2020; Post et al., 2011), Sustainalytics (Kristianti, 2023), Crisil (Singh et al., 2024), Sino-Securities ESG Rating System (Yan et al., 2024), Rankings (RKS) rating (Lyu et al., 2022; McGuinness et al., 2017; Naveed et al., 2021), People and Planet (Aly et al., 2024), IOW/future ranking (Dienes and Velte, 2016), CSRHub (Nerantzidis et al., 2022), China Research Service Platform (Peng and Chandarasupsang, 2023), SynTao Green Finance (Xiaomei et al., 2021) or China Securities Index Co. Ltd. (Fan et al., 2024).

Chapter 4

Data

4.1 Identification of Primary Studies

To analyze publication bias in the literature that estimates the effect of female directors on ESG performance, we compile an original dataset. Through a quantitative research design, 547 effect estimates from 108 unique research papers are collected. These so-called primary studies are identified using the Google Scholar search engine. To ensure that the search query is robust enough, we refine it iteratively until the five most essential studies appear among the top search results (Iršová et al., 2023). The reason why Google Scholar is strongly preferred is its ability to search the full text of papers, which allows the inclusion of studies that may lack all relevant keywords in their title or abstract (Gechert et al., 2022). In order to make the search process replicable, we stick to a single search query. Thus, the final search query used for the identification of the studies included in this meta-analysis has the following form:

Board Gender Diversity ESG (”*Disclosure*” *OR* ”*Score*” *OR* ”*Performance*”)

The search performed in May 2024 returned more than 22,800 results. We examined the abstracts and results sections of the first 500 studies to determine the presence of the desired estimates. Subsequently, employing the snowballing method, references of the primary studies were downloaded and sorted by the frequency of citation using Scopus. The first 100 of them were reviewed. In addition, the citations of the three most important primary studies in the last three years were inspected. To ensure that we potentially do not exclude any recent study, we repeat the Google Scholar search and limit the search to past three

years. The first 30 results are inspected. Shortly before finalizing the dataset, we perform the search again in December 2024 and check first 50 results in order not to exclude any study that has been published since May 2024. The search identifies eight new studies.

The studies in our final sample satisfy the following criteria. First, studies examine the impact of female leadership on ESG performance. Second, studies employ the three pillars of ESG performance, namely the pillars of environmental, social, and corporate governance. Third, selected studies report uncertainty measures such as standard errors, confidence intervals or p-values. Due to the high number of identified studies eligible for inclusion, we further restrict our inclusion criteria. Thus, elaborating on the first inclusion criteria, only studies that employ board gender diversity as a fraction of women over all board members are considered. As a result, studies that use statistical measures to quantify the degree of diversity within a group (such as the Blau or Shannon index (Abdullah et al., 2024; Gangi et al., 2021) and studies that focus solely on female CEO or chair leadership (Aabo and Giorici, 2023; Harjoto and Wang, 2020; Li et al., 2023; Liu et al., 2024), proportion of women in top management teams (Fu et al., 2023), dummies for female board representation (Chebbi and Ammer, 2022) or report only number of women on board (Kravchenko et al., 2023) are not included.

Moreover, we strictly adhere to research papers that have their ESG performance rated by Bloomberg or LSEG. Thus, in relation to the response variable, we disregard studies that use ratings from other sources. The reason is that most of them report ESG scores on different scales, which would force us to use partial correlation coefficients, leading to a loss of economic interpretation.

Furthermore, we do not use studies that employ authors' own ESG metrics (even though based on GRI or SASB). From our perspective, self–constructed ESG scores based on checklists can be subjective and influenced by the researchers' own preferences. In contrast, Bloomberg and LSEG ratings are reported on a scale of 0.01 to 100 and Bloomberg's ratings resemble quite well those LSEG ones (Dorfleitner et al., 2015).

There exist a handful of studies that are not considered in our final sample despite the use of Bloomberg or LSEG data. Usually, these employ logarithmic transformation of the ESG score or discount the original ESG score by the controversy score (Shakil, 2021), which represents the frequency and severity of the negative ESG scandals reported in the media. Similarly, these are also excluded for comparability reasons.

To finalize the criteria, two practical obstacles prevented us from incorporating

Table 4.1: Primary studies used

Adamu et al. (2024)	de Klerk & Singh (2023)	Moussa & Elmarzouky (2023)
Adeneye et al. (2024)	De Masi et al. (2020)	Nadeem et al. (2017)
Adiasih & Lianawati (2018)	Dicuonzo et al. (2024)	Nandi et al. (2023)
Agnese et al. (2024a)	Disli et al. (2021)	Nekhili et a. (2021)
Agnese et al. (2024b)	Donkor et al. (2023)	Nery & Morales (2022)
Agustina & Barukah (2024)	Ellili (2022)	Nicolo et al. (2021)
Ahmadi & Amara (2024)	Fahad & Rahman (2020)	Nicolo et al. (2023)
Al Kurdi et al. (2023)	Gaio & Goncalves (2022)	Nicolo et al. (2024)
Al-Shaer et al. (2024)	Gerged et al. (2023)	Nuhu et al. (2024)
Ali & Firmansyah (2023)	Giannarakis (2013)	Ozturk (2023)
Aliani et al. (2024)	Govindan et al. (2021)	Paolone et al. (2024a)
Aliti & Nguyen (2023)	Grubler (2024)	Paolone et al. (2024b)
Alkayed et al. (2024)	Gungor & Seker (2022)	Pinheiro et al. (2024)
Alkhawaja et al. (2023)	Halid et al. (2022)	Pinheiro et al. (2023)
Almaqtari et al. (2023)	Heubeck (2024)	Pirskanen (2023)
Almaqtari et al. (2024)	Husted & de Sousa-Filho (2019)	Qureshi et al. (2019)
Amara & Ahmadi (2024)	Chebbi et al. (2020)	Qureshi et al. (2022)
Amorelli & Garcia-Sanchez (2023)	Issa et al. (2022)	Rella & L'Abate (2022)
Andreassen & Bukhari (2024)	Jizi et al. (2022)	Sari & Fitriani (2023)
Arayssi et al. (2016)	Kamaludin et al. (2022)	Setiani & Novitasari (2024)
Arayssi et al. (2020)	Kamran et al. (2023)	Shahbaz et al. (2020)
Arayssi et al. (2024)	Kampoowale et al. (2024)	Shakil et al. (2021)
Arduino et al. (2024)	Khatri (2023)	Sofiati & Mita (2024)
Ben Fatma & Chouaibi (2021)	Khemakhem et al. (2023)	Temiz & Acar (2023)
Benaguid et al. (2023)	Kouki (2023)	Toerien et al. (2023)
Bhatia & Marwaha (2022)	Lavin & Montecinos-Pearce (2021)	Trireksani et al. (2024)
Bigelli et al. (2023)	Lozano & Martinez-Ferrero (2022)	Uyar et al. (2020)
Birindelli et al. (2018)	Makeeva et al. (2022)	Uyar et al. (2021)
Boukattaya & Omri (2021)	Manita et al. (2018)	Van Hoang et al. (2023)
Bruna et al. (2021)	Marrone et al. (2024)	van Zundert (2024)
Buallay et al. (2022)	Martinez et al. (2020)	Velte (2016)
Cucari et al. (2018)	Martinez et al. (2022)	Wang et al. (2022)
Dakhli (2021)	Meen (2023)	Waterstraat et al. (2021)
Dang et al. (2021)	Mehmood et al. (2023)	Wu et al. (2024)
Dang et al. (2023a)	Miranda et al. (2023)	Yadav & Prashar (2022)
Dang et al. (2023b)	Monteiro et al. (2024)	Yarram & Adapa (2021)

Notes: The table represent the list of primary studies identified through Google Scholar search engine. The last study was added on December 16, 2024.

all existing research papers on this topic. It is mainly accessibility issues that lead us to exclude studies that are not available through the standard university license, even though these studies have the potential to provide the desired estimates. The complete process of study identification is represented in Appendix A Figure A.1.

Following Stanley (2001), we exclude no study based on its publication form. Consequently, the final sample includes not only standard research articles from peer-reviewed journals but also working papers, articles and master theses. We account for the quality of the study in our analysis. Table 4.1 presents the final list of primary studies used in our meta-analysis.

4.2 Dataset Compilation

The individual effect estimates and their uncertainty measures from the primary studies in our list are systematically collected together with various other variables to examine the systematic heterogeneity among the reported coefficients. These variables encompass estimation characteristics, specific data features, dummy variables for control variables, and publication characteristics.

Upon completion of data collection, we proceed with some data adjustments to ensure that our dataset contains comparable effect estimates. First, some studies explore a non-linear relationship between female leadership and ESG performance by including a quadratic term (Birindelli et al., 2018). To handle the presence of two related estimates, we follow the methodology of Žigraiova and Havránek (2016) and linearize the effect as follows:

$$\beta_{is} = \hat{\beta}_{lis} + 2\hat{\beta}_{qis}\overline{x}_{es}, \tag{4.1}$$

$$SE(\beta_{is}) = \sqrt{SE(\hat{\beta}_{lis})^2 + 4SE(\hat{\beta}_{qis})^2\overline{x}_{es}^2} \tag{4.2}$$

where $\hat{\beta}_{lis}$ denotes the linear estimate of the female leadership coefficient for the i-th estimation in study s, $\hat{\beta}_{qis}$ refers to the estimate of the quadratic term, and $\overline{x}_{es}$ is the sample mean of female leadership reported in study s. In addition, $SE(\hat{\beta}_{lis})$ and $SE(\hat{\beta}_{qis})$ represent the standard errors for the linear and quadratic term estimates, respectively.

Second, some studies in our dataset model interaction terms between female leadership and other variables, such as the common law tradition (Alkhawaja et al., 2023), the critical mass of women on the board (Birindelli et al., 2018), or ESG controversies (Shakil et al., 2021). In line with the approach by Cazachevici et al. (2020), we compute the average marginal effect of female leadership on firm performance, applying the delta method to derive the corresponding standard errors:

$$\beta_{is} = \hat{\beta}_{lis} + \hat{\beta}_{tis}\overline{x}_{es}, \tag{4.3}$$

$$SE(\beta_{is}) = \sqrt{SE(\hat{\beta}_{lis})^2 + SE(\hat{\beta}_{tis})^2\overline{x}_{es}^2} \tag{4.4}$$

where β_{is} denotes the calculated marginal effects of female leadership, $\hat{\beta}_{lis}$ represents the estimated linear effect size of female leadership, and $\hat{\beta}_{tis}$ corresponds

to the estimate of the interaction term in the i-th estimation of the study s. The calculation also considers the mean value of the variable included in the interaction term $\overline{x}_{es}$. The standard errors are obtained using the reported standard errors of the linear term, $SE(\hat{\beta}_{lis})$, and the interaction term, $SE(\hat{\beta}_{tis})$. It is important to highlight that for some studies, the lack of summary statistics or measures of uncertainty for quadratic or interaction terms prevents us from applying this transformation. In cases where the mean value for the variable included in the interaction term is too high and produces extremely large average marginal effects, we exclude such observations.

Similarly, after the use of the delta method for the recalculation of normalized variables in the study of Giannarakis (2013), extremely large estimates are produced. Thus, we also disregard these observations.

Finally, it is essential to note that the authors use ESG scores and board gender diversity differently. The vast majority either use combination of raw ESG scores (ranging between 0.01 and 100) and percentage of women on boards (0-100%), or normalize both to a 0–1 scale. However, some studies normalize female leadership to the 0–1 scale while using original ratings. To ensure consistent interpretation, estimates and standard errors are re-scaled in such cases.

The final phase of data transformation involves several adjustments. For example, Bruna et al. (2021) report standard errors of zero for certain estimates. To tackle this issue, we replace standard errors of 0.000 with 0.0001, ensuring a sufficiently small value.

For studies reporting 5% significance level, we estimate the t-statistic as 2.27 (Fahad and Rahman, 2020; Kamaludin et al., 2022). Similarly, for studies that either report a significance level of 1% (Miranda et al., 2023; Nuhu and Alam, 2024) or only report p-values itself as 0.000 (Bhatia and Marwaha, 2022; Nadeem et al., 2017; Setiani and Novitasari, 2024) we set p-values to 0.0001. For studies such as (Martínez et al., 2022), which systematically report only zero p-values for all estimates of interaction terms of interest, we approximate p-values by random values between 0.0000 and 0.0004.

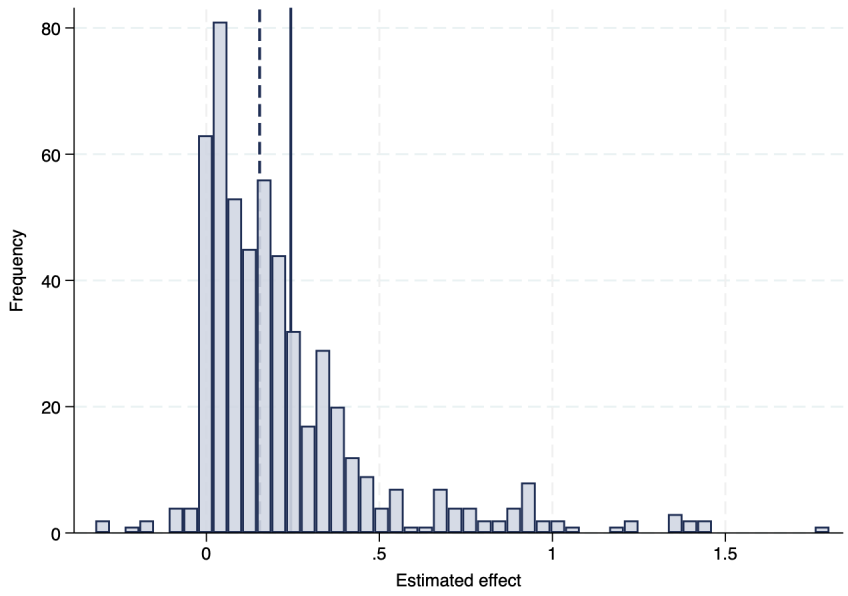
Finally, despite rigorous data cleaning, some outliers in the estimated size effects, along with their associated standard errors, remain. To retain the informational content of these observations without skewing the results, we winsorize these extreme values and uncertainty measures at the 1% level.

4.3 Descriptive Evidence

First, it is essential to present summary statistics to offer a detailed overview of the dataset before analyzing publication bias and heterogeneity. Following a thorough selection of relevant studies and the necessary data adjustments, the dataset includes 547 estimates derived from 108 primary studies investigating the effect of female leaders on corporate ESG performance. The studies included in the sample span between years 2014 and 2024, which aligns with the introduction of ESG ratings. Furthermore, the fact that only 16 studies were published before 2021 shows increasing interest in this topic. In addition, as pointed out by Husted and de Sousa-Filho (2019), researchers may have been waiting for the time when companies have more women on their boards. Obviously, the effect of female directors on the sustainability performance might not be possible to study when there are no such directors at all.

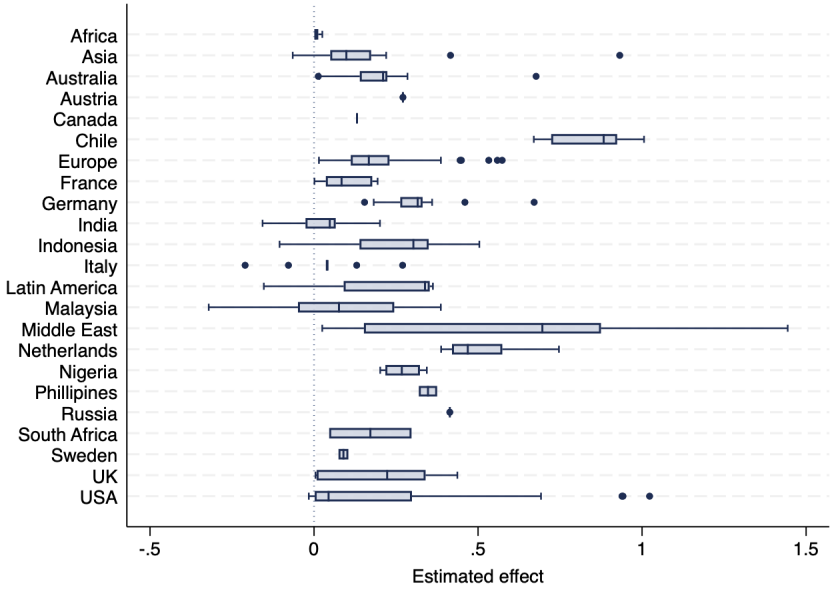
Figure 4.1 portraits a first insight into the distribution of estimated effects. Encompassing the estimates within the interquartile range only, histogram shows a heavy-tailed distribution with peak around zero and positive skewness.

Figure 4.1: Distribution of gender diversity effects



Notes: The figure presents a distribution of the estimated effects of board gender diversity on ESG performance as reported across individual studies. The mean reported effect, represented by a solid vertical line, indicates a 0.244 point increase in ESG rating following a 1% increase in female board representation. The distribution of estimates is shown within the interquartile range. The dashed vertical line represents the median.

Figure 4.2: Variation of estimates within and across countries



Notes: The figure shows a box plot of the estimated effects. The inner box line indicates median. The whiskers extend to the most extreme values within 1.5 times the interquartile range. Extreme outliers are omitted from the figure but remain included in all statistical tests.

Similarly, the box plot depicted in Figure 4.2 illustrates the heterogeneity of estimates across different countries. Consistent with the histogram presented in Figure 4.1, most of the estimates are positive. Similarly, some distributions show more symmetry, while others are more skewed. Interestlingy, some regions show median values closer to zero, while others, on average, exhibit higher estimated effect. There are obviously some regions, such as Latin America, Malaysia or Middle East, that exhibit larger variation with wider interquartile ranges and longer whiskers.

However, to gain some initial insights into the heterogeneity, it is important to examine the mean values of the estimated effects. Table 4.2 reports both weighted and unweighted mean values for different categories. Weighted means are preferred due to the fact that in our dataset there are primary studies that report disproportionally many estimates (Alkhawaja et al., 2023) compared to few other primary studies that report only one estimate (Waterstraat et al., 2021; Yarram and Adapa, 2021). Although the overall sample means do not differ dramatically, some subsample means vary much more. The overall estimated effect of female board participation on firm's ESG is around 0.248 (weighted) and 0.242 (unweighted). Both means suggest a slight but positive relationship between

board gender diversity and ESG scores.

Looking closer at the mean values for different categories, studies that employ panel data, which accounts for time variation, yield more conservative estimates. Following the same logic, one could expect that employing more rigorous estimation techniques should produce less biased and thus lower estimates. However, it is not the case on the basis of the reported results. Contrary, among studies that do not employ proper endogeneity control the effect seems to be more pronounced. The same holds for regions with low board gender diversity, such as the Middle East, where the effect of women on ESG performance is much stronger than in European or US countries. Based on previous research (Berg et al., 2022; Dorfleitner et al., 2015), one could also expect that using ESG scores from different data providers might produce differences in the mean values of the estimated effects. In this case, simple means show slightly higher estimated effect for sub-samples employing LSEG data.

At this stage, determining whether there is publication bias remains a challenge. Although the weighted mean indicates substantial differences in the estimated effect between published and unpublished studies, simple averages can obscure the underlying distortions in the reported findings (Ioannidis et al., 2017). To ensure a more comprehensive understanding, the following chapter investigates in more detail whether empirical evidence on female leadership and ESG performance is subject to publication bias.

Table 4.2: Board gender diversity effects in different contexts

		Weighted			Unweighted		
	Est.	Mean	95% conf. int.		Mean	95% conf. int.	
All estimates	547	0.248	0.216	0.279	0.242	0.214	0.271
Data type							
Cross-sectional data	21	0.514	0.221	0.807	0.349	0.037	0.661
Panel data	526	0.228	0.199	0.258	0.238	0.211	0.265
ESG data provider							
Bloomberg data	205	0.228	0.169	0.288	0.255	0.201	0.309
LSEG data	342	0.260	0.223	0.296	0.235	0.203	0.266
Publication status							
Published	520	0.258	0.224	0.291	0.243	0.214	0.272
Unpublished	27	0.160	0.076	0.243	0.230	0.134	0.326
Estimation techniques							
Linear estimation	189	0.232	0.184	0.280	0.219	0.176	0.263
Panel method	217	0.223	0.181	0.264	0.234	0.193	0.276
Instrumental variable	37	0.378	0.182	0.574	0.308	0.154	0.463
GMM	47	0.320	0.163	0.477	0.245	0.116	0.374
Other technique	57	0.218	0.133	0.303	0.304	0.213	0.395
Endogeneity							
Poor endog. control	442	0.264	0.227	0.300	0.251	0.219	0.282
Proper endog. control	105	0.191	0.130	0.252	0.207	0.141	0.274
Spatial variation							
Global	178	0.263	0.199	0.327	0.202	0.149	0.255
Europe	182	0.183	0.162	0.203	0.199	0.177	0.221
USA	55	0.199	0.131	0.267	0.198	0.126	0.270
Asia	43	0.165	0.043	0.287	0.169	0.059	0.278
Middle East	30	0.577	0.394	0.759	0.657	0.459	0.854
Other countries	59	0.451	0.283	0.619	0.382	0.269	0.494
Economic Development							
Developed	287	0.231	0.194	0.267	0.215	0.188	0.241
Emerging	99	0.266	0.182	0.350	0.403	0.310	0.497
Mixed	161	0.276	0.200	0.351	0.192	0.134	0.250
Sector							
Financial	37	0.290	0.200	0.380	0.382	0.260	0.504
Non-financial	152	0.244	0.191	0.297	0.221	0.184	0.259
Mixed	358	0.243	0.200	0.286	0.237	0.199	0.275
Estimate type							
Preferred estimate	454	0.245	0.209	0.281	0.233	0.203	0.263
Others	93	0.265	0.199	0.331	0.287	0.210	0.364

Notes: The table displays summary statistics of the estimated effect of board gender diversity on ESG outcomes across various subgroups within the literature. "Est." = number of estimates. The right-hand panel (unweighted) reports simple means along with their associated 95% confidence intervals, treating each estimate equally. In the left-hand panel (weighted) estimates are weighted by the inverse of the number of estimates reported per study, assigning equal weight to each study.

Chapter 5

Unveiling Publication Bias

The concept of publication bias can be interpreted in several ways. Broadly defined, it refers to the divergence between the originally estimated effect size and the estimate that ultimately gets published. More narrowly, it refers to the correlation between the estimated and their standard errors. This correlation is mainly driven by the general preference of researchers and journals to prioritize studies with estimates of intuitive sign and statistical relevance. Thus, estimated effects are often unnaturally inflated to the point in which standard errors offset required statistical significance.

Consequently, three key mechanisms contributing to publication bias are described (Card and Krueger, 1995). It all starts with the bias among editors and reviewers toward studies that align with the prevailing theoretical perspectives. Then it spreads to researches who strategically adjust the model specifications. The circle closes with an overall preference across the academic community for statistically significant findings. Publication bias is, however, not necessarily the result of intentional manipulation. Researchers may be inclined to submit significant findings under the assumption that such evidence is more likely to be published, while it might be also prioritized by editors and referees who eventually believe to provide compellingly informative results. Regardless of the underlying motivations, this dynamic can lead to systematic underreporting of nonsignificant results, ultimately biasing the body of available research (Stanley, 2005).

While the issue of publication bias has been widely documented across various economic research fields such as minimum wage effects (Card and Krueger, 1995; Doucouliagos and Stanley, 2009), fiscal policy (Heinemann et al., 2018) or behavioral economics (Havránek et al., 2017), we believe that the empirical literature on the relationship between female leadership and ESG performance may

be another candidate. It seems to be particularly vulnerable for several reasons. First, the topic is closely aligned with appealing narratives advocating for gender equality. In addition, based on women's nurturing and empathetic potential, reporting nonsignificant female effect on sustainable business practices would not be so compelling. As a result, studies that show a positive and statistically significant relationship between board gender diversity and ESG outcomes may be receiving more attention. If publication bias is indeed present, the descriptive statistics reported in the previous chapter may not accurately reflect the true underlying effect. A biased sample of estimates would distort the average reported relationship, potentially overstating the strength or consistency of the effect of female leadership on ESG performance. Therefore, a more rigorous assessment is required to evaluate the credibility of the existing evidence and to test for the presence of publication selection mechanisms in the literature.

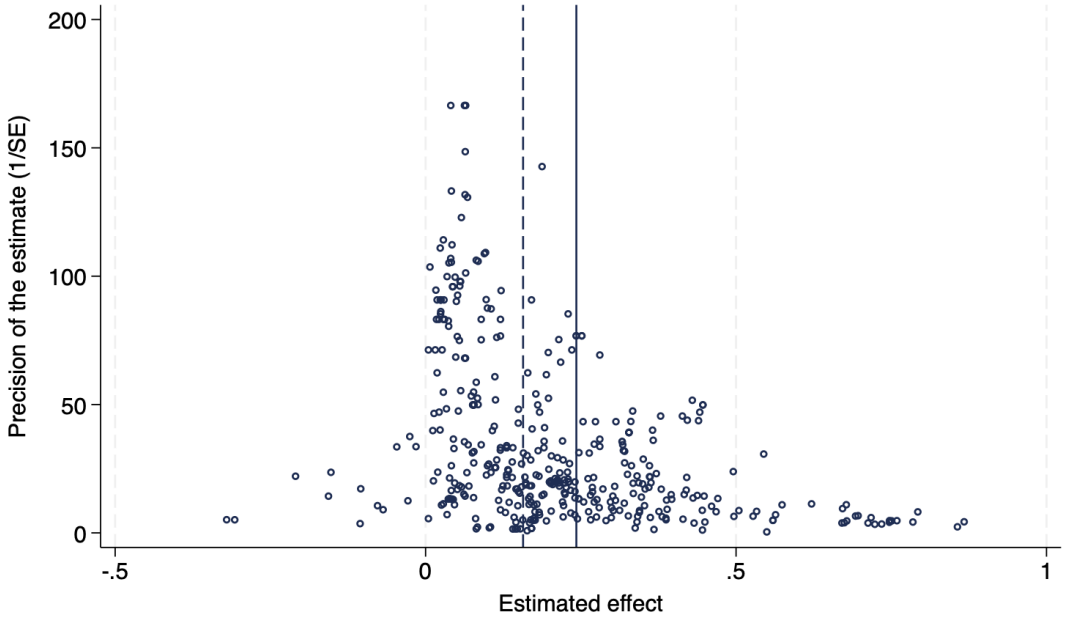
5.1 Graphical Method

To initiate the assessment of potential publication selection bias, we employ a widely used visual diagnostic tool for meta-analyses, the funnel plot (Egger et al., 1997). This plot is a scatter diagram in which individual effect estimates are displayed along the horizontal axis, while their corresponding precision, commonly measured as the inverse of the standard error, is plotted on the vertical axis.

Under the assumption of no publication bias, the most precise estimates are expected to cluster at the top of the graph around the average effect, forming a symmetrically inverted funnel shape (Stanley, 2005). The spread of points should widen progressively toward the bottom, where less precise estimates are located. This pattern indicates that the distribution of the findings reflects random variation around the size of the true effect (Sterne and Harbord, 2004).

Applying funnel plot to our sample of estimates produces a slightly skewed inverted funnel shape depicted in the Figure 5.1. Those most precise estimates are unanimously clustered left to the mean effect, signaling the presence of publication bias. However, it is important to interpret the funnel plot asymmetry with caution, as it can also be driven by other factors such as heterogeneity between studies, differences in methodology or data quality issues (Egger et al., 1997). Additionally, such analysis has a subjective nature as it strongly depends on the researchers, which presentation of the funnel plot is provided. For example, our funnel plot has been cut from above and positive values larger than 1 have been removed. Nevertheless, we analyzed various forms of funnel plot and each of

Figure 5.1: Funnel plot of reported estimates



Notes: In the absence of publication bias, the most precise estimates are expected to cluster around the mean effect. Less precise estimates should be symmetrically distributed around it. The figure suggests an asymmetry with the most precise estimates reported between zero and mean effect, denoted by the solid vertical line. The dashed line represents the median effect. Extreme outliers are omitted from the figure but remain included in all statistical tests.

them shared two common characteristics - negative estimates are slightly under-represented and those most precise estimates are located between zero and sample mean of estimated effect. Although this graphical method suggests a right-skewed funnel plot, we strengthen the validity of our visual inspection in the following section by complementing the funnel plot analysis with formal statistical tests for asymmetry.

5.2 Linear Methods

Card and Krueger (1995) in their study propose a linear approach to test for publication bias, estimating the following regression:

$$E_{is} = \beta_0 + \beta_1 SE_{E_{is}} + \epsilon_{is}, \tag{5.1}$$

where E_{is} are estimated effects and $SE_{E_{is}}$ are their corresponding standard errors, ϵ_{is} represents error term, β_0 is effect beyond bias and β_1 reflects the publication bias.

Regression denoted in Equation 5.1, also called Egger regression, measures the symmetry of the funnel plot. As the regression suggests, the correlation between reported estimates and their standard errors will be examined. If publication bias is present, the correlation will be significantly different from zero (Stanley, 2005). Transformed into hypothesis testing, this would test the null hypothesis of publication bias being absent against the alternative hypothesis of publication bias being present.

Based on the classification of Doucouliagos and Stanley (2013), if the funnel asymmetry test (FAT) is not statistically significant or the absolute value of β_1 is less than 1, the degree of publication bias is classified as little to modest. Publication bias is considered substantial if FAT is significant and the absolute value of β_1 ranges between 1 and 2. Consequently, if β_1 exceeds 2 and FAT is significant, publication bias is classified as severe.

Equation 5.1 can be estimated using various model specifications. We begin with ordinary least squares. Second, we estimate model that uses the between-study variance. Conversely, model that employs within-study variance is not estimated as some primary studies provide only one effect estimate. Third, we employ weighting schemes commonly applied in meta-analyses, e.g. Gechert et al. (2022) or Havránek et al. (2018) - weighting by the inverse of the number of observation per study and by precision of estimates. The first weighting ensures that studies with different sample sizes are given equal weight, while the latter one ensures that less weight is given to less precise estimates and the heteroscedasticity from FAT is eliminated. This is achieved by dividing each term by their standard errors:

$$E_{is}\frac{1}{SE_{E_{is}}} = t_{is} = \beta_0\frac{1}{SE_{E_{is}}} + \beta_1 SE_{E_{is}} + \vartheta_{is}, \tag{5.2}$$

where β_0 denotes the direction and magnitude of the potential publication bias, while β_1 signifies the average effect adjusted for this bias, t_{is} is the t-statistic of the i-th estimated effect of study s. Equation β_1 denotes precision estimate test (PET) (Stanley, 2005).

As a methodological refinement, we address the potential heteroskedasticity in the FAT-PET framework by clustering standard errors at the study level. The standard assumption of independently and identically distributed error terms $\epsilon_{is} \sim iid$ is likely violated due to within-study correlations among reported estimates. Clustering standard errors at the study level allows us to account for this intra-study dependence while maintaining the assumption of independence

Table 5.1: Linear tests

	OLS	Between Effects	Study Weight	Precision Weight	IV
Publication Bias	0.886***	0.887***	0.965**	2.283***	2.113***
(*Standard Error*)	(0.309)	(0.151)	(0.411)	(0.405)	(0.669)
	[0.266, 1.896]		[0.041, 1.95]	[1.416, 3.177]	{0.444, 4.365}
Mean Beyond Bias	0.162***	0.154***	0.192***	0.035*	0.051
(*Constant*)	(0.025)	(0.032)	(0.04)	(0.021)	(0.061)
	[0.105, 0.217]		[0.109, 0.278]	[0.004, 0.125]	{ -0.068, 0.17}
First-stage robust F-stat					11.53
Observations	547	547	547	547	547
Studies	108	108	108	108	108

Notes: The results of the regression $E_{is} = \beta_0 + \beta_1 \cdot \text{SE}(E_{is}) + \varepsilon_{is}$, where E_{is} denotes the i-th estimated effect in the study s, and $\text{SE}(E_{is})$ denotes its standard error. 95% confidence intervals from wild bootstrap (Roodman et al., 2019) are shown in square brackets. For IV, the Anderson-Rubin 95% confidence interval recommended by Anderson and Rubin (1949) is shown in curly brackets. Standard errors and confidence intervals are enclosed in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

across different studies. Although our sample meets the minimum threshold for valid inference, the presence of unequal cluster sizes may still introduce bias, as noted by MacKinnon and Webb (2017). To mitigate this issue, we adopt the wild cluster bootstrap method proposed by Gechert et al. (2019), which is particularly robust to cluster imbalance. We report 95% confidence intervals based on this procedure for all specifications, with the exception of the between-effects estimation at the study level.

Finally, because estimating the Egger regression might suffer from endogeneity between estimates and their standard errors, Stanley (2005) recommends the usage of the instrumental variable method. Bias usually arises from random sampling errors and joint computation of estimates and their standard errors, both potentially influenced by the estimation method chosen in the primary study. An instrument, which would satisfy both the relevance and exogeneity assumptions, is the square root of the number of observations. Studies that use larger samples should produce smaller standard errors and sample sizes should not be correlated with a chosen estimation method.

Table 5.1 summarizes the results from the model specifications described above. All show positive and significant publication bias at the 5% level. Although the first three specifications are categorized as mild, the remaining two show a severe level of publication bias, according to the classification suggested by Doucouliagos and Stanley (2013). Compared to the weighted and unweighted means from the primary studies (0.248 and 0.242, see Table 4.2), the bias-corrected mean is notably lower, ranging from 0.035 to 0.192.

5.3 Non-linear Methods

Funnel asymmetry and precision-effect tests (FAT-PET) are widely used to detect publication bias and to estimate the true underlying effect (Stanley, 2008). However, a key limitation of these approaches lies in their assumption of a linear relationship between standard errors and estimates. This assumption may not hold in practice, especially for highly precise estimates. Their inherently small standard errors can yield significance even without selective reporting (Stanley et al., 2010). Thus, they are less likely to be influenced by publication bias. As a result, linear models may exaggerate the extent of bias and underestimate the true effect. To address this issue, we complement the linear FAT-PET approach with a set of non-linear estimators.

As a first step in our non-linear checks, we implement the weighted average of adequately powered (WAAP) estimator, introduced by Ioannidis et al. (2017). This approach aims to mitigate the impact of publication bias by focusing solely on studies with statistical power exceeding 80%. Hence, smaller and underpowered studies that could potentially inflate bias are excluded. The adequately powered estimates are then aggregated using optimal inverse-variance weights ($1/SE^2$) to derive a more reliable average effect size. However, a known limitation of WAAP, as noted by Stanley et al. (2017), is its dependency on the presence of sufficiently powered studies within the dataset. In cases where such studies are scarce or absent, WAAP becomes uninformative. Thus, we complement it with a few other checks.

Selection model due to Andrews and Kasy (2019) builds on the assumption

Table 5.2: Non-Linear tests

	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		0.012		7.905*	
		(0.009)		(4.210)	
Effect Beyond Bias	0.022***	0.098	0.066***	0.020***	0.133***
	(0.005)	(0.086)	(0.006)	(0.002)	(0.037)
# of estimates			155		
% of information			42%		
Observations	547	547	547	547	547

Notes: WAAP denotes weighted average of adequately powered estimates (Ioannidis et al. (2017)); Selection model denotes technique due Andrews and Kasy (2019); Stem denotes the stem-based technique (Furukawa (2019)); Kink denotes the endogenous kink model (Bom and Rachinger (2019)); p-uniform* denotes the technique due to van Aert and van Assen (2021). Standard errors and confidence intervals are enclosed in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

that the likelihood of publishing an effect estimate is influenced by its statistical significance. Specifically, the model posits that this probability shifts once the estimate surpasses certain t-statistic thresholds. Using maximum likelihood estimation, the model calculates the publication probabilities for different data windows defined by these critical t-statistic thresholds. Then it identifies which estimates are underrepresented within these windows and assigns them greater weight, thus adjusting for potential publication bias. This approach is similar in spirit to the one proposed by Hedges (1992), but provides a more nuanced treatment of the relationship between statistical significance and the likelihood of publication.

Third, two non-linear methods are applied. The STEM-based method and the Endogenous Kink method, both of which extend the logic behind the "Top 10" approach introduced by Stanley et al. (2010). These approaches assume that a subsample of the most precise estimates is less prone to publication bias, thus providing a more accurate estimate of the true underlying effect. The STEM-based method, developed by Furukawa (2019), focuses on the "stem" of the funnel plot, selecting the most precise estimates by minimizing the mean squared error. The Endogenous Kink method due to Bom and Rachinger (2019) fits a piecewise linear meta-regression of estimates against their standard errors. The regression consists of two segments. The flat segment represents the most precise estimates unaffected by standard errors and a positively sloped segment indicates the correlation between standard errors and estimates impacted by publication bias. The intersection of these two segments, known as the "kink", serves as the threshold for identifying the most reliable estimates. Both methods endogenously determine the proportion of precise estimates to include in the final calculation of the average effect. These should be large enough to ensure efficiency, but small enough to exclude imprecise estimates. Improving upon the "Top 10" approach and thus avoiding the arbitrary cut-off, both methods aim at providing a more efficient estimate of the true effect.

To account for a possible endogenous relationship between estimates and their standard errors, the p-uniform* method developed by van Aert and van Assen (2021) is employed. In our case, the p-uniform* is estimated using the method of moments, which is based on a principle that p-values should ideally follow a uniform distribution. However, publication bias distorts this distribution. Specifically, estimates with p-values just below the significance threshold tend to be over-represented, while those with p-values just above the 5% level are underrepresented. The objective of the p-uniform* method is to identify a value around

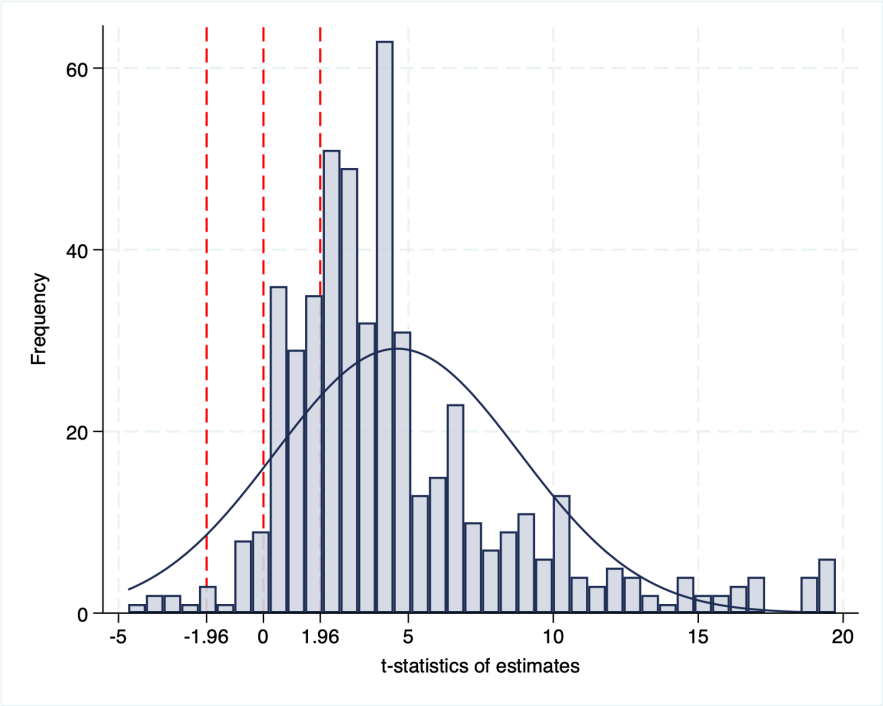
which the p-values align with a uniform distribution.

The results of above described non-linear tests are presented in Table 5.2. On average, the estimated effects beyond bias generated by non-linear tests are slightly lower than the ones generated by linear tests, ranging from 0.02 to 0.133.

5.4 Caliper Test

To conclude our investigation into publication bias, which has been covered in linear tests using the instrumental variable method and by p-uniform* in nonlinear tests, the Caliper test is applied. In this method proposed by Gerber et al. (2008), t-statistics are grouped into equal-sized ranges called "calipers" around critical thresholds such as -1.96 or 1.96. Then, the number of estimates in each range is compared. Under the assumption of no publication bias, the estimates just below and just above these thresholds are expected to appear with equal frequency. Figure 5.2, which shows the distribution of t-statistics for the collected effects of board gender diversity on ESG performance, different scenario can be

Figure 5.2: Distribution of t-statistics of estimates



Notes: The figure presents a distribution of t-statistics. Dashed vertical lines denote the critical thresholds of -1.96, 0, and 1.96. Extreme outliers are omitted from the figure but remain included in all statistical tests.

inferred. Looking at the 0 and 1.96 thresholds, there are significant irregularities in the distribution. The share of observations just above both of the thresholds is much greater than below. Due to the fact that around -1.96 there are almost no observations, we apply Caliper tests only at 0 and 1.96 thresholds. The results are presented in Table 5.3. It must be noted that the Caliper widths are a bit wider than usual because of the small number of observations within the required ranges. However, results for Caliper widths of 0.35 and 0.45 show clear publication selection bias between significant and non-significant positive estimates. For the critical threshold of t-statistic 1.96, positive significant estimates make up 64% of observations within the 35% interval, while positive non-significant estimates account for 36%. This finding strengthens our analysis by indicating that researchers not only favor positive over negative estimates, as supported by linear and non-linear tests, but also show a preference for statistically significant positive results over non-significant ones.

Table 5.3: Caliper test

Critical threshold of t-statistic	0	1.96
Caliper size: 35%	0.9^{***}	0.644^{***}
Share of estimates above the threshold	(0.069)	(0.072)
	N=20	N=45
Caliper size: 45%	0.76^{***}	0.645^{***}
Share of estimates above the threshold	(0.087)	(0.061)
	N=25	N=62

Notes: The reported shares are statistically significantly above the 0.5 threshold. Standard errors in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.5 Robustness Checks

To provide robust evidence on publication bias, the testing is repeated on a distinct subsample of the original dataset. Figure and tables related to robustness check are presented in Appendix B.

The subset utilized for robustness check discards estimates that are derived from the interaction term between female leadership and other variables, or whose uncertainty measures are adjusted by some approximation as described in Chapter 4. Demonstrated by descriptive statistics in Table 4.2, such estimates yield higher weighted mean compared to direct estimates. In case of interaction terms,

applying the delta method to compute the average marginal effect and the corresponding standard errors includes the mean value of the other variable. In case its average is disproportionally large, it may be reflected in the inflated corresponding average marginal effect, as well as the underlying standard error. Based on our observations, there were no variables included in the interaction terms that would, contrariwise, have too small means. Similarly, imputing p-values or approximating standard errors might introduce a systemic error in standard errors. Thus, a subset of 454 estimates is created and the presence of publication bias is checked.

The subsample check is robust and uniform in the reported results. These resemble quite precisely the results of the full-sample testing, with publication bias being present among all linear tests. Similarly to the main results, the Selection model in Table B.2 does not confirm significant publication bias. When it comes to the true effect, magnitudes do not diverge dramatically. Although both linear and non-linear analyses reject the null hypothesis of no effect, the bias-adjusted mean remains close to zero, suggesting a negligible economic impact. Since this result holds consistently, we conclude that while female board participation has a positive effect on sustainable business practices, the final magnitude is rather small. If ESG scores are measured on a scale from 0.01 to 100, a 1% increase in female board representation yields an average effect size of 0.02 to 0.13.

Chapter 6

Drivers of Heterogeneity

Heterogeneity in the literature may influence the correlation between estimates and standard errors, originally attributed to publication bias. Therefore, it is necessary to test whether the results of publication bias will remain robust to inclusion of control variables reflecting the study design. Simultaneously, we identify those that systematically affect heterogeneity among reported estimates. According to Adams et al. (2015), these could be the use of different samples, time windows or empirical methods. To explore the sources of heterogeneity among the reported estimates, we characterize following variables further organized into five categories: data characteristics, publication characteristics, estimation techniques, design of the analysis and control variables.

6.1 Description of Effect Drivers

Data Characteristics

First, we record whether the study uses *Panel data*, captured by a dummy variable. In contrast to cross-sectional data, longitudinal datasets allow for the control of time-invariant unobserved heterogeneity.

Second, we include a dummy variable indicating whether the study relies on ESG data from *LSEG*, to distinguish the estimates obtained from studies using Bloomberg's ESG data. Previous research has documented substantial divergence in ESG scores between data providers, ultimately affecting empirical results (Dorfleitner et al., 2015).

Third, the variable *Average data year* denotes the standardized average year of the data used in each study. It serves as a control for potential unobserved year-fixed effects, such as the mainstreaming of ESG practices, which may in

turn reduce the marginal effect of female leadership, as well as temporal shifts in methodology of ESG data or evolving regulatory requirements.

Lastly, we construct the variable *Board gender diversity*. This variable captures the average representation of women on boards in the data used for the estimation. Dang et al. (2021) suggests that the effect of female directors on CSR is non-linear and likely depends on surpassing a certain threshold. Usually, this represents exactly one female director on the board, in line with the token theory due to Kanter (2008). When a single woman is appointed to a board, she can be perceived as a token member (Torchia et al., 2011), without the opportunity to offer a different perspective in boardroom discussions. Recent studies suggest that the influence of women on the board is more likely to be significant when there is a critical mass of at least three women is present (Konrad et al., 2008; Torchia et al., 2011). By including this variable, we account for heterogeneity in baseline diversity between study samples, a factor that could meaningfully influence the estimated effect.

Publication Characteristics

Next, several characteristics related to publication are employed to control for possible sources of reporting or publication bias. First, we introduce a dummy variable for estimates produced by studies which got *Published* in a peer-reviewed publication. It is well documented that studies published in academic journals can systematically differ from those in working papers or reports due to selective reporting, editorial standards, and the peer review process, which tends to favor statistically significant results.

Second, we account for the visibility and academic influence of each study using the logarithm of the total number of *Citations* a primary study has received.

Third, we collect the Recursive Impact Factor (RIF) from IDEAS/RePEc (Research Papers in Economics), introduced by Zimmermann (2013), which provides a measure of journal quality based on the citation frequency of articles published in a journal. Studies not listed or unpublished studies are assigned an *Impact factor* of zero. The RIF serves as a more relevant measure for studies in fields such as management, sustainability and business where journals may not be represented in traditional citation indices such as the Journal Citation Reports (JCR).

Estimation Techniques

We categorize the estimation methods used to investigate the relationship into four distinct approaches: *Linear estimation*, which includes ordinary least squares (OLS) and generalized least squares (GLS) models; the *Panel method*, if fixed effect (FE) or random effects (RE) models are used; the *Instrumental variable* technique, which includes two-stage least squares (2SLS), control function or limited information maximum likelihood (LIML) models; and finally, the *GMM method*. Each of these methods operates under different assumptions, with varying degrees of ability to account for unobserved differences. In the article of Dang et al. (2023b), various estimation techniques are employed on a single dataset and reported estimates vary significantly in their magnitude.

Design of the Analysis

We incorporate several variables that relate to the design of the analysis, which may influence the reported estimates due to varying methodological approaches and sample characteristics.

First, we classify studies based on their control for endogeneity. As highlighted by Dang et al. (2023a), the relationship between board gender diversity and ESG performance suffers from three sources of endogeneity - omitted or unobserved firm characteristics, reverse causality and dynamic endogeneity. To account for this, we categorize the studies into three levels: *Poor endogeneity control*, if authors control for none or maximum one endogeneity type and *Good endogeneity control* if authors control for two to all endogeneity sources. This framework allows us to capture the varying degree to which studies address potential biases in their estimates.

Second, we classify the primary studies based on the geographical scope of their datasets. We distinguish between *Europe*, the *USA*, *Asia*, and the *Middle East*, using *Other countries* as the reference group. In addition, we code a separate category, *Global*, to account for studies relying on multinational datasets, typically combining data from major economies like the USA, the EU, China and other. These studies cannot be meaningfully assigned to any single region. Including them in the reference group would risk diluting region-specific effects. The distinction between regions is important, as societal norms and regulatory frameworks vary substantially, which in turn affects the estimated relationships.

We also introduce a variable, *Emerging markets*, to account for studies conducted in emerging market contexts.

Table 6.1: Overview and descriptive statistics of contextual variables

Variable	Description	Mean	SD	WM
Estimate	= estimated effect	0.243	0.337	0.248
Standard error	= estimated standard error of the estimate	0.091	0.156	0.107
Data characteristics				
Panel data	= 1 if panel data is used for estimation	0.962	0.192	0.932
LSEG data	= 1 if LSEG's ESG data is employed	0.625	0.485	0.620
Average data year	= logarithm of the average data year	7.608	0.001	7.608
Average sample BGD	= logarithm of the average sample board gender diversity	2.809	0.650	2.799
Publication characteristics				
Published	= 1 if study is published in a peer reviewed journal	0.951	0.217	0.898
Citations	= logarithm of total citations	2.821	1.555	2.915
Impact factor	= logarithm of journal's impact factor	-2.403	2.009	-2.411
Estimation techniques				
Linear estimation	= 1 if OLS or GLS is used	0.346	0.476	0.380
Panel method	= 1 if FE or RE is used	0.397	0.490	0.378
Instrumental variable	= 1 if 2SLS, CF or LIML is used	0.068	0.251	0.057
GMM	= 1 if GMM or its extension is used	0.086	0.281	0.132
Other technique	= 1 if other estimation technique is used (reference category for estimation technique)	0.104	0.306	0.053
Design of the analysis				
Poor endog. control	= 1 if poor endogeneity control is employed	0.808	0.394	0.783
Proper endog. control	= 1 if proper endogeneity control is employed (reference category for endogeneity control)	0.191	0.394	0.217
Global	= 1 if study employs a global sample of firms	0.325	0.469	0.222
Europe	= 1 if study focuses on European firms	0.333	0.472	0.364
USA	= 1 if study focuses on US firms	0.101	0.301	0.128
Asia	= 1 if study focuses on Asian firms	0.079	0.269	0.135
Middle East	= 1 if study focuses on firms in the Middle East	0.055	0.228	0.056
Other countries	= 1 if study focuses on firms from other countries	0.107	0.310	0.095
Emerging markets	= 1 if study focuses on firms in emerging markets	0.181	0.385	0.261
Financial sector	= 1 if study focuses on financial firms	0.068	0.251	0.097
Number of firms	= logarithm of number of firms	5.723	1.751	5.348
Control variables				
Firm size	= 1 if study controls for size of the firm	0.892	0.310	0.880
Board independence	= 1 if study controls for independence of the board	0.678	0.467	0.688
CSR committee	= 1 if study controls for existence of firm's corporate social responsibility committee	0.378	0.485	0.363

Notes: SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study, LSEG = London Stock Exchange Group, BGD = board gender diversity, CF = control function, LIML = limited information maximum likelihood, GMM = Generalized Method of Moments, CSR = corporate social responsibility

As discussed by Yilmaz et al. (2021), companies in emerging economies may have different levels of non-financial performance, which could be the result of distinct institutional frameworks, legal systems or political environments.

Next, we include a variable *Financial sector* to distinguish studies focusing on financial institutions. Based on Paolone et al. (2024), the effect of female leadership on ESG performance among financial institutions often exhibits different patterns due to the strict regulatory environment in which such companies operate. This in many countries affects not only the incorporation of sustainable practices, but also the appointment of women to boards.

Due to high correlation coefficients of *Sample size* with other variables, we decided to employ a proxy, which should equally well explain the differences in estimates. *Number of firms* represents the logarithmic transformation of the total number of companies included in the analysis. Smaller samples may fail to accurately capture the true effect of female leadership on corporate outcomes.

Control Variables

In our analysis, we include several control variables to account for factors that may influence the relationship studied. Firstly, *Firm size* should control for distinct governance dynamics of larger companies with complex organizational structures and more resources at hand. Similarly, *Board Independence* captures the potential effect of independent directors who are more likely to advocate for non-financial goals such as environmental and social responsibilities. Independent directors are less beholden to the management and may respond more precisely to external pressures. Finally, the existence of a *CSR committee* demonstrates the commitment and orientation of a board towards responsible behavior, signaling the firm's efforts to invest in stronger stakeholder relations (Hussain et al., 2018). The inclusion of this variable could absorb some of the effects of board gender diversity, as CSR committees typically oversee and guide the firm's CSR activities, potentially overlapping with the influence of female leadership on board behavior.

6.2 Estimation

Once baseline categories are excluded, 24 explanatory variables remain. While each could influence the reported board gender diversity effects, only a few are likely to matter consistently. This creates considerable model uncertainty. Adding all variables into one model would likely produce very imprecise estimates, even

for the key ones. Thus, an intuitive approach does not come at hand. It would introduce inflated standard errors due to a high number of predictors. In contrast, selecting a single *best* model among all possible combinations would not only be arbitrary, but it would also ignore the uncertainty that inherently comes with such a decision.

That is where Bayesian model averaging (BMA) enters the picture. Instead of relying on a single model specification, BMA treats the model space itself as uncertain and builds in a probabilistic framework to address this. It does so by evaluating all possible combinations of explanatory variables, each of them forming its own potential model, and assigning to each one a posterior model probability (PMP) based on how well it explains the data (Raftery et al., 1995). This means that instead of choosing one model and discarding the rest, all models are weighted by how likely they are to be the *true* model. The result is a robust and transparent approach that acknowledges the inherent uncertainty in model selection and uses it to our advantage (Fragkiskos et al., 2023; Havránek et al., 2015).

Formally, in the context of Bayesian model averaging, let Y be an outcome variable with K potential explanatory variables $x = \{X_k\}_{k=1}^{K}$ and a sample of observations on Y and x to be denoted as $D = \{y_i, x_{1i}, x_{2i}, \ldots, x_{Ki}\}_{i=1}^{n}$. Each model M_j corresponds to a unique subset $x_j \subseteq x$ of the explanatory variables, resulting in a model space of 2^K possible models. We denote this space as $\{M_j\}_{j=1}^{2^K}$, where k indexes the explanatory variables and j indexes the models.

The posterior probability of a specific model M_j given observed data D is derived from Bayes' theorem:

$$p(M_j|D) = \frac{p(D|M_j)p(M_j)}{p(D)} = \frac{p(D|M_j)p(M_j)}{\sum_{l=1}^{2^K} p(D|M_l)p(M_l)}, \tag{6.1}$$

where $p(D|M_j)$ is the marginal likelihood of the data under model M_j, which reflects the model's ability to explain the observed data, and $p(M_j)$ is the prior probability assigned to model M_j, capturing beliefs about the model's plausibility before observing data (Zeugner, 2011). The denominator sums over the entire model space, integrating out model uncertainty to ensure proper normalization.

Consequently, if $\hat{\beta}_{kj}$ is the estimated coefficient for variable X_k in model M_j and $E(\beta_k|D, M_j) = \hat{\beta}_{kj}$, then a model-weighted average of the posterior means of a regression coefficient β_k, weighted by the posterior probability of each model, is given by:

$$E(\beta_k|D) = \sum_{j=1}^{2^K} E(\beta_k|D, M_j) \cdot p(M_j|D) = \sum_{j=1}^{2^K} \hat{\beta}_{kj} \cdot p(M_j|D), \tag{6.2}$$

Its associated weighted posterior variance (or posterior standard deviation) follows naturally:

$$Var(\beta_k|D) = \sum_{j=1}^{2^K} \left(Var(\beta_k|D, M_j) + \hat{\beta}_{kj}^2\right) p(M_j|D) - E(\beta_k|D)^2, \tag{6.3}$$

A key strength of BMA is its ability to quantify the relative importance of each variable through the posterior inclusion probability (PIP). This statistic indicates the probability that a given variable truly belongs in the model, after accounting for the evidence across the entire model space:

$$PIP_k = \sum_{j=1}^{2^K} p(M_j \mid \beta_{kj} \neq 0, D). \tag{6.4}$$

PIP sums PMPs of all models that include the k-th variable. Its value ranges from 0 to 1. If PIP=1, then k-th variable appears in each effective model. Following a scale proposed by Kass and Raftery (1995), the effect of variables may be categorized as weak, positive, strong or decisive, based on PIP value falling into intervals of 0.5-0.75, 0.75.-0.95, 0.95-0.99 and 0.99-1 respectively.

After having introduced BMA method, following meta-regression equation may be estimated:

$$EE_{is} = \beta_0 + \beta_1 X_{is} + \beta_2 SE_{EE_{is}} + \epsilon_{is} \tag{6.5}$$

where EE_{is} represents the estimated effect, X_{is} is an additional predictor, and $SE_{EE_{is}}$ is standard error. While β_0 is a constant, β_1 represent direction and intensity of publication bias and classically, $\epsilon_{i,t}$ refers to the error term.

The meta-regression is estimated using the *bms* package in R, which applies the Markov chain Monte Carlo approach via the Metropolis-Hastings algorithm to avoid the infeasibility of computing all 2^{24} possible models. Instead of evaluating each model, the algorithm explores those with the highest posterior inclusion probabilities by iteratively comparing and selecting between models based on their relative posterior support (Zeugner, 2011). To account for the uneven number of estimates reported per study, the data is weighted by the number of observations.

Our baseline BMA estimation adopts the unit information prior (UIP), which

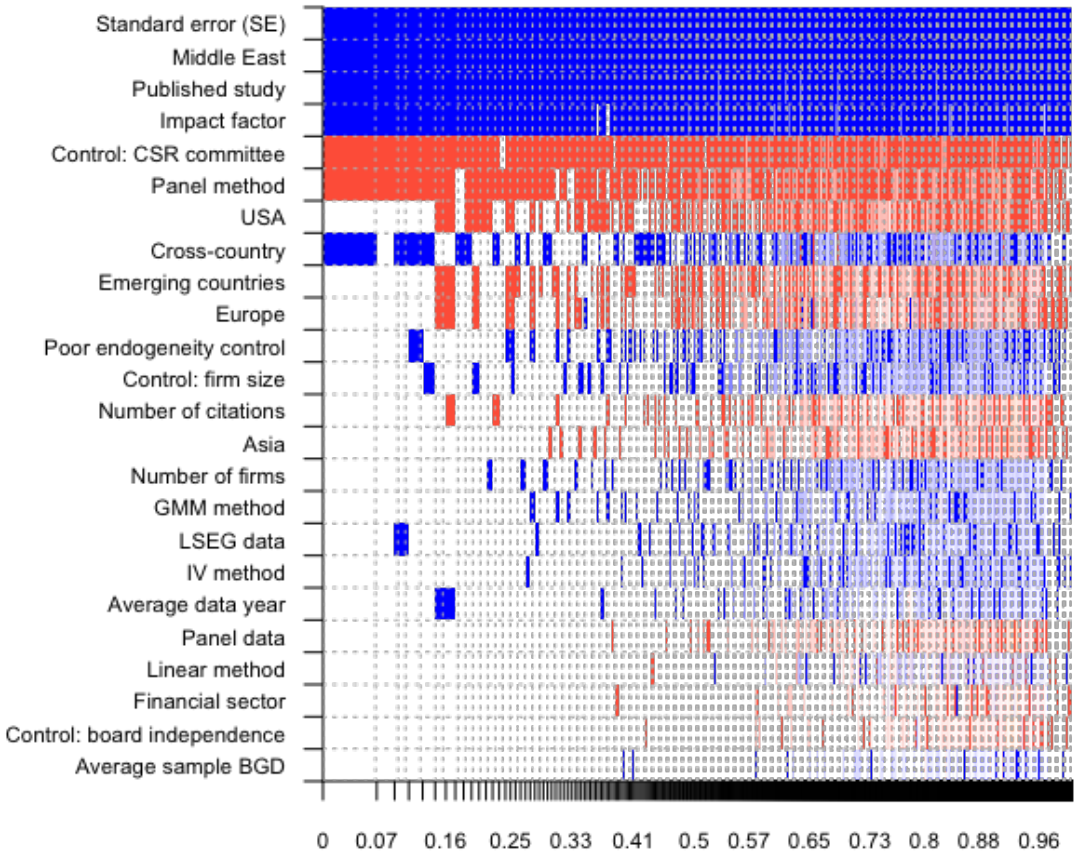
assumes that the prior carries the same weight as one observation (Eicher et al., 2011). This is due to the limited prior information we have at our disposal. To address collinearity, we follow George (2010) and apply dilution prior. This prior gives less weight to models with a small determinant of the correlation matrix (Hasan et al., 2018).

In addition to the baseline BMA model, we introduce multiple robustness checks. Firstly, a variation of BMA is employed. Instead of using UIP, we try Bayesian risk information criterion (BRIC) g-prior due to Fernández et al. (2001) and the beta-binominal model prior according to Ley and Steel (2009). Second, Frequentist model averaging (FMA) and frequentist check are employed. While Bayesian model averaging brings prior beliefs into the equation, Frequentist Model Averaging (FMA) keeps it strictly empirical, drawing solely from the data at hand. This makes FMA an attractive complement to BMA, especially when aiming for a more objective assessment. To put this into practice, we follow the well-established path laid out by Havránek et al. (2017), who propose using the Mallows model averaging estimator. Here, model weights are assigned by minimizing the so-called Mallows criterion, a function that rewards models with lower variance and better in-sample fit. This technique, inspired by the work of Magnus et al. (2010) and refined through orthogonalization of the covariate space (Amini and Parmeter, 2012), helps to considerably narrow down the model space, especially useful when working with higher number of explanatory variables. The implementation is kept fully transparent, building on openly shared code (Havránek et al., 2021), which allows for consistent replication. As a frequentist robustness check, Equation 6.5 is re-estimated using OLS with study-level clustered standard errors. Again, only variables surpassing the 0.5 threshold in posterior inclusion probability from the baseline BMA are included, ensuring that the analysis centers on those with the most empirical support.

6.3 Results

The results of the baseline BMA estimation are visualized in Figure 6.1. The vertical axis ranks the explanatory variables based on their posterior inclusion probabilities, while the horizontal axis represents the cumulative posterior model probability. In other words, variables at the top of the figure are the ones that best explain the estimated effects of board gender diversity on ESG performance and the models on the left offer the optimal balance between goodness of fit and parsimony. Blue color (darker in grayscale) indicates positive estimate parameter

Figure 6.1: Model inclusion in Bayesian model averaging



Notes: The figure shows results of the baseline BMA estimation reported in Table 6.2 based on best 10587 models (g-prior = UIP, model prior = dilution). The vertical axis ranks the explanatory variables based on their posterior inclusion probabilities, from highest at the top to lowest at the bottom. The horizontal axis represents the cumulative posterior model probability. A blue (darker) color indicates positive effect, red (lighter) color negative effect and absence of color denotes non-inclusion of the corresponding variable in the respective model. Variable descriptions in Table 6.1 and further technical diagnostic information in Appendix C Table C.2 and Figure C.3.

of the corresponding predictor. For example, the color is universally blue for *Middle East*, which suggests that studies focusing on firms in the Middle East typically estimate more positive effects. Red color (lighter in grayscale) means the estimated parameter of the corresponding explanatory variable is negative. Blank cells mean that the corresponding explanatory variable is excluded from the model. The figure makes it clear that most of the 24 variables do not help explain the systematic differences in reported effect of female leadership. Only 6 variables are robustly important, and their corresponding regression coefficients have the same sign irrespective of other controls being included or ignored.

Table 6.2: Results of baseline BMA and frequentist check estimations

	Bayesian Model Averaging (baseline model)			**Stepwise regression (frequentist check)**		
	P. Mean	*P. SD*	*PIP*	*Coef.*	*SE*	*p-value*
Standard error (SE)	0.900	0.075	1.000	0.876	0.069	0.000
Data characteristics						
Panel data	-0.008	0.034	0.077			
LSEG	0.006	0.019	0.111			
Average data year	0.003	0.009	0.088			
Average sample BGD	0.001	0.006	0.027			
Publication characteristics						
Published	0.247	0.064	0.988	0.189	0.031	0.000
Citations	-0.003	0.008	0.164			
Impact factor	0.025	0.009	0.944	-0.004	-0.002	0.021
Estimation techniques						
Linear estimation	0.000	0.013	0.048			
Panel method	-0.071	0.047	0.774	-0.079	0.028	0.005
Instrumental variable	0.010	0.036	0.101			
GMM	0.015	0.051	0.115			
Design of the analysis						
Poor endog. control	0.023	0.050	0.228			
Global	0.040	0.057	0.416			
Europe	-0.031	0.060	0.271			
USA	-0.065	0.086	0.448			
Asia	-0.015	0.042	0.152			
Middle East	0.383	0.078	0.999	0.350	0.063	0.000
Emerging markets	-0.048	0.080	0.329			
Financial sector	-0.001	0.010	0.034			
Number of firms	0.003	0.008	0.145			
Control variables						
Firm size	0.019	0.045	0.184			
Board independence	-0.001	0.008	0.034			
CSR committee	-0.096	0.042	0.914	-0.086	0.029	0.004
Studies	108			108		
Observations	547			547		

Notes: The table displays results for baseline BMA model and frequentist check that includes only variables with PIP > 0.5 in the BMA baseline model. Baseline BMA uses the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively. Variables are described in Table 6.1. P. Mean = Posterior Mean. P. SD = Posterior Standard Deviation. PIP = Posterior Inclusion Probability.

Quantitative Bayesian model averaging outputs reported in Table 6.2 refer to the visual evidence provided in Figure 6.1. Similarly, in Appendix C the robustness checks of the baseline BMA model are provided. On the left-hand side of Table C.1 we report BMA that employs different priors, while on the right-hand side of the same table we provide FMA results. In BMA estimations, the marginal effect of a given study characteristic is represented by the posterior mean, while its effectiveness in explaining the differences in the estimated effects reported in the literature is represented by the posterior inclusion probability. For example, when study controls for CSR committee the estimated board gender diversity premium typically decreases by almost 0.1 points compared to studies that omit this control. This variable also exhibits a high posterior inclusion probability of 91%. This suggests its high relevance. However, given that employed ESG ratings are measured on a scale of 0.01 to 100, the average effect change is not so large. Yet, what drives a slightly larger effect change is the geographical location of firms employed for analysis. All BMA specifications, including the baseline and robustness checks, consistently show that studies focusing on firms in the Middle East report substantially higher board gender diversity premiums. To understand this result, we examine the common characteristics of the underlying observations. Notably, all studies that cover Middle Eastern countries feature very low average female board representation, typically between 2 and 5%. At first glance, one might expect that such a low initial representation could amplify the estimated effect of a marginal increase in gender diversity. However, the reality is more complex. From a theoretical standpoint, such small numbers often correspond to tokenism (Kanter, 2008), where female board members lack the critical mass necessary to influence firm strategy or governance, particularly in contexts where gender norms may be restrictive. This raises doubts about the interpretation of the estimated effect as a genuine causal relationship. Importantly, the variable capturing sample board gender diversity does not significantly explain heterogeneity in the results. This suggests that the larger reported effects for the Middle East may reflect other factors. For example, it is possible that the few firms in the region appointing women to their boards are not representative. They may already excel in ESG performance due to international ownership, progressive values or external pressure. Alternatively, cultural or publication dynamics may incentivize selective reporting of positive effects. Taken together, while the pattern is robust across models, its interpretation should be approached with caution.

Although both BMA specifications yield negative coefficient estimates for regions such as *Europe*, *USA*, and *Asia*, the corresponding posterior inclusion prob-

abilities remain slightly below the conventional threshold of 50%. Nonetheless, Frequentist model averaging provides complementary evidence, identifying these regional indicators as significant drivers that systematically lower the reported board gender diversity premium relative to the reference group. The reason why other regions do not survive the BMA estimation could be that the Middle East samples, where large effects are consistently found, may suppress the estimated effect in other regions. In other words, European or US samples do not explain the heterogeneity of estimated effects that precisely as Middle Eastern ones. While *Emerging* and *Global* ar reported as significant in FMA robustness estimation, *Panel method* becomes irrelevant predictor of the gender diversity there. Contrary to our expectations regarding the direction of the estimated coefficient for studies focusing on firms in emerging markets, the variable *Emerging* consistently yields a negative coefficient across all model specifications. However, upon closer inspection of the data, a straightforward explanation emerges. A substantial number of observations stem from countries such as Indonesia, Malaysia, and India, which are also captured by the broader *Asia* variable, likely driving the observed result.

Nonetheless, it is the standard error that appears to carry the greatest weight in driving the heterogeneity. This predictor is associated with the largest increase in the estimated effect of female leadership. We interpret this as robust evidence of publication bias, even after accounting for heterogeneity.

All in all, the analysis of model uncertainty yields two main takeaways. First, publication bias remains robust even with the inclusion of explanatory variables. In fact, most of the effect beyond publication bias is driven by the correlation between board gender diversity and standard error. Second, next to standard error, only five additional variables out of 24 cross the PIP threshold of 0.5. More specifically, *Middle East*, *Published study*, *Impact factor*, *Control: CSR committee* and *Panel method* yield significant changes in the estimated effects of female leadership.

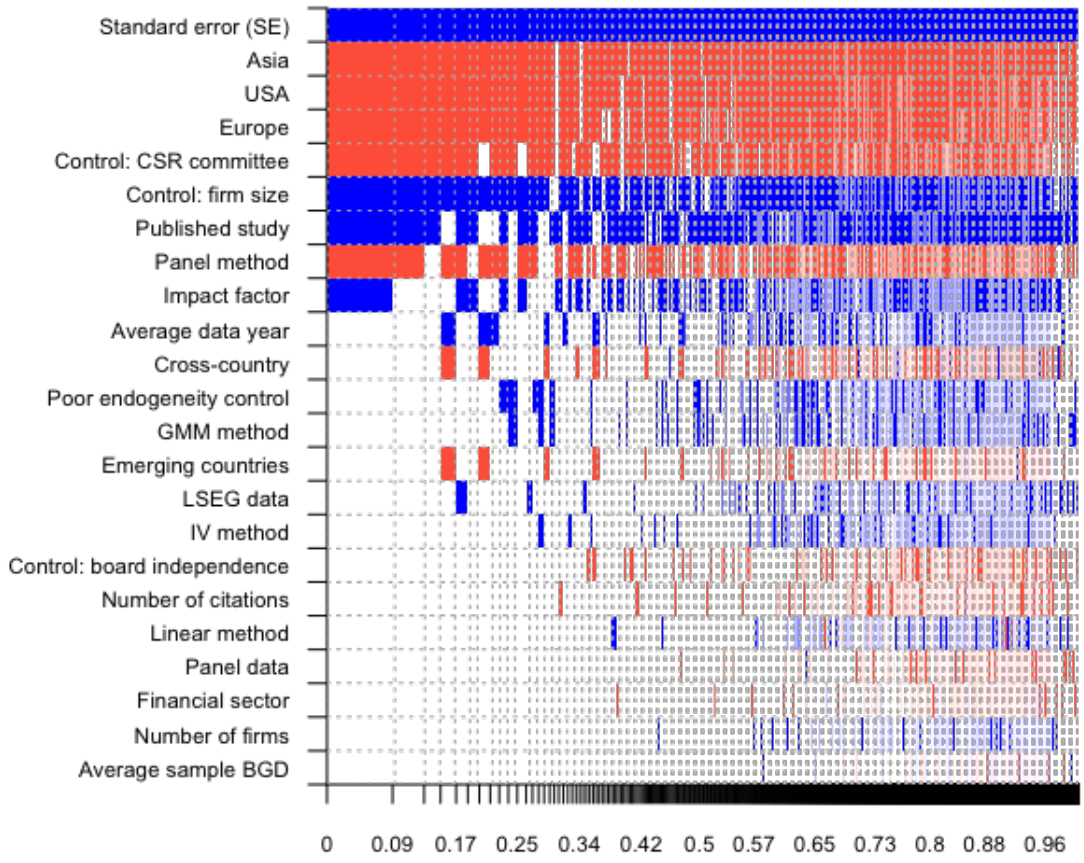
To be precise, *Impact factor* fails to pass the frequentist check reported in the right-hand part of Table 6.2. Due to the fact that many studies are published in journals without any ranking, imputing zeros for these observations could introduce some systematic error. The remaining variables included in the analysis do not exhibit a systematic association with the reported effects. Surprisingly, it does not seem to matter whether linear or instrumental variable approach is employed, or whether researchers properly control for endogeneity. Similarly, no evidence is found on the ESG data used or the sector analyzed.

6.4 Subsample Robustness Check

Given concerns about the particular interpretation of the *Middle East* category, we conduct a robustness check on a subsample of the original data, excluding observations from this region. As already suggested by Frequentist Model Averaging, there might be also other regions driving the heterogeneity in a significant way.

The results of this check are presented in Appendix C. Looking at Figure 6.2 and reported inclusion probabilities in Table C.4, we observe four new predictors of effect heterogeneity once we exclude primary studies that focus solely on the

Figure 6.2: Model inclusion in subsample robustness BMA



Notes: The figure shows results of the subsample robustness BMA estimation reported in Table C.4 based on best 10729 models (g-prior = UIP, model prior = dilution). The vertical axis ranks the explanatory variables based on their posterior inclusion probabilities, from highest at the top to lowest at the bottom. The horizontal axis represents the cumulative posterior model probability. A blue (darker) color indicates positive effect, red (lighter) color negative effect and absence of color denotes non-inclusion of the corresponding variable in the respective model. Variable descriptions in Table 6.1 and further technical diagnostic information in Appendix C Table C.5 and Figure C.5.

Middle East. Namely, *Europe*, *USA*, *Asia* and *Firm size* enter the game. For the questionable variable *Impact factor* the subsample robustness BMA finds an inclusion probability slightly below 50%, excluding it from the frequentist check. Both the subsample robustness BMA and frequentist check confirm our concerns about the importance of other regions, which has been suppressed by *Middle East*. However, in contrast to the Middle East, studies analyzing European, US and Asian firms estimate the female leadership effect to be lower compared to studies analysing firms in other countries. Additionally, after dropping all Middle East observations, *Firm size* becomes an important predictor of the estimated effect. Including a control for firm size typically enhances the gender diversity premium by approximately 0.12 points compared to the primary studies which omit this control. This finding may suggest that the true impact of board gender diversity is underestimated when studies do not control for firm size, as the confounding influence of firm size dilutes the observed relationship.

6.5 The Best Practice Estimate

The findings of the previous analyses indicate that the literature examining the ESG performance implications of female leadership suffers from positive publication bias. Furthermore, the heterogeneity in collected estimates can be explained by specific study characteristics. In this section, we explore what the mean effect would be if we controlled for these influences. In particular, using the synthetic study approach, we construct a "best practice" estimate. Practically, the best practice estimate is obtained by estimating a linear regression where each covariate is set to a preferred value. Variables reflecting poor research practices are assigned their sample minimum. Variables representing best practices are assigned their sample maximum. In cases where no preference is expressed, we use the sample mean.

While this procedure provides a useful benchmark, it inevitably involves subjective judgment. From the outset, it is important to acknowledge that defining a best practice estimate is inherently normative, as it reflects decisions about the characteristics of an ideal study. In our view, an ideal estimate should satisfy several conditions. First, it should be derived using panel data, which enables researchers to better distinguish causal effects from correlations. Hence, we set the use of *Panel data* to its sample maximum. Second, despite BMA findings suggesting estimation methods are not systematically influential, we favor techniques that mitigate endogeneity bias. Therefore, we assign maximum values to

Table 6.3: Predicted best practice estimates

	Best practice estimate	95% Confidence Interval
All	0.109	(-0.257 ; 0.475)
Europe	0.107	(-0.284 ; 0.497)
USA	0.065	(-0.327 ; 0.458)
Asia	0.015	(-0.392 ; 0.422)
Middle East	0.123	(-0.347 ; 0.593)
Global	0.038	(-0.309 ; 0.386)
LSEG	0.124	(-0.232 ; 0.481)
Bloomberg	0.083	(-0.306 ; 0.472)

Notes: The table reports the "best practice" estimate of the board gender diversity premium in different contexts. The rationale behind the "best practice" selections is outlined in the text. Confidence intervals are calculated using OLS with standard errors clustered at the study level.

Instrumental variable and *GMM* estimators, while setting zero to *Poor endogeneity control*. Third, as we aim to approximate the most recent evidence, we set the *Average data year* to its sample maximum. Moreover, we prioritize estimates from published studies in journals with high *Impact factors*, believing that they are more likely to adhere to rigorous academic standards. Consequently, we set this publication-related characteristic at its sample maxima. For the *Number of citations*, we use the sample mean, as citation counts do not necessarily correlate with study quality. To minimize the influence of publication bias, we set the standard error to its sample minimum. As larger sample sizes yield more precise estimates, *Number of firms*, employed as a proxy for sample size, is set to the maximum value. Additionally, we assign a value of 1 to controls for *Firm size* and *CSR committee* membership, as these variables are considered crucial for accurately modeling ESG performance. For all other variables, we retain the sample mean without imposing further assumptions.

Following these preferences, Table 6.3 reports the corresponding best practice estimates along with their 95% confidence intervals. Confidence intervals are constructed using OLS estimation with standard errors clustered at the study level.

Predicted best practice estimates reported in Table 6.3 summarize the expected board gender diversity premium under preferred study conditions. The overall best practice estimate equals 0.109. When examining differences across regions, we observe only modest variation. Studies focusing on firms in the Middle East

report the highest premium (0.123), followed closely by those based on European firms (0.107). The USA and Asia exhibit lower best practice estimates of 0.065 and 0.015, respectively, while studies employing global samples report a premium of 0.038. Turning to the employed ESG data sources, studies relying on LSEG data suggest a premium of 0.124, whereas studies based on Bloomberg data yield a slightly lower estimate of 0.083.

However, it is important to note that all implied best practice estimates are surrounded by wide confidence intervals, all of which include zero. This indicates substantial uncertainty regarding the precise magnitude of the effect. Even when varying the weighting scheme, the size of the predicted effect remains small and statistically insignificant. To sum up, these results align with the findings from previous sections. While the direction of the board gender diversity premium remains positive across most contexts, its magnitude appears limited, suggesting that the economic relevance of the relationship is likely modest.

Chapter 7

Conclusion

The objective of this thesis was to conduct a quantitative review of studies investigating the impact of female leadership on ESG outcomes. Specifically, we tested for the presence of publication bias using graphical and statistical methods. Furthermore, we identified study characteristics explaining heterogeneity among reported estimates using Bayesian averaging model, an estimation method accounting for model uncertainty. To carry out these analyses, we constructed an original dataset comprising 547 estimates drawn from 108 primary studies.

From a theoretical perspective, this work contributes to the existing literature in three ways. First, to the best of our knowledge, this thesis represents the first meta-analysis that synthesizes evidence on the relationship between board gender diversity and ESG ratings, utilizing modern quantitative techniques. Second, by pooling results from a large number of studies, we systematically test for the presence of publication selection bias in the literature. Finally, we explore which study characteristics, methodological aspects and control variables drive the heterogeneity in the reported estimates.

The empirical results allow us to confirm all three of our main hypotheses. First, we find visual and statistical evidence for positive publication bias. This suggests that studies reporting positive effects of board gender diversity on ESG outcomes may have a higher probability of being published. Second, the presence of publication bias exaggerates the mean effect reported in the literature. Consequently, the corrected mean effect is negligible, ranging between 0.02 and 0.192 points across linear and non-linear tests that incorporate various weighting schemes and specifications. Apparently, while gender diversity may contribute positively to ESG performance, it is not a dominant factor. The low economic magnitude of the corrected mean reinforces this interpretation - a firm would need

a 10% increase in board gender diversity to experience a shift in the ESG score of 2 to 19,2 points retrospectively. Third, our BMA results highlight that the geographical focus of the primary studies is the most systematic driver of heterogeneity among reported estimates. This is in line with meta-analysis of Wu et al. (2022) studying the effect of board gender diversity on disclosure of corporate social performance. In particular, our results suggest that studies focusing on firms in the Middle East report substantially larger board gender diversity premiums across all model averaging specifications. As this variable is found as second most important in explaining the underlying heterogeneity of the estimates after standard error, we suspect that it could suppress the effect of other potentially important study characteristics. Thus, we conduct a robustness check excluding studies on Middle Eastern firms. Based on this restricted sample, we find that estimates for Europe, the USA, and Asia are significantly lower compared to the reference category.

In addition, we note a few other interesting observations. While the selection of ESG data does not appear to play a role, controlling for CSR committee seems to absorb some of the board gender diversity premium. Furthermore, once the Middle East observations are excluded, studies controlling for firm size are found to report larger effects of female leadership on firms' ESG performance.

Nevertheless, the limitations of this study should be acknowledged. Due to a huge number of studies identified, we restricted our sample to studies using Bloomberg or LSEG ESG data only. This brings two main implications. First, we potentially omit important variation in studies employing other metrics of ESG performance. Second, it indirectly limits our analysis to listed companies. These are the ones for which Bloomberg and LSEG published the ESG ratings. Thus, future research could extend the analysis to private companies, incorporating studies utilizing also other metrics of ESG performance to assess whether the observed patterns persist. Additionally, should we have had a larger sample of observations, it would be insightful to explore whether the effect of female leadership on ESG outcomes varies among industries where female representation in top management is more pronounced, such as in healthcare or education, compared to sectors like IT or energy, where gender diversity is often lower (Deloitte Insights, 2023; Csonka and Milhomem, 2023; World Economic Forum, 2024). Investigating such sectoral differences could provide further nuances to our understanding of how female leadership interacts with ESG performance.

To sum up, while our findings provide evidence of a positive association between board gender diversity and ESG outcomes, the true magnitude of the rela-

tionship appears modest once publication bias and study heterogeneity are taken into account. Policymakers, investors and researchers alike should maintain cautious expectations regarding the impact of female leadership on corporate sustainability performance. Board gender diversity should be regarded as a contributory factor to ESG performance, but cannot be considered as its robust standalone predictor. Initiatives to promote female leadership should be accompanied by broader structural changes, such as the establishment of CSR committees, if the goal is to materially impact ESG outcomes.

Bibliography

Aabo, T. and Giorici, I. C. (2023). Do female CEOs matter for ESG scores? *Global Finance Journal*, 56:100722.

Abdul Rahman, R. and Alsayegh, M. F. (2021). Determinants of corporate environment, social and governance (ESG) reporting among Asian firms. *Journal of Risk and Financial Management*, 14(4):167.

Abdullah, A., Yamak, S., Korzhenitskaya, A., Rahimi, R., and McClellan, J. (2024). Sustainable development: The role of sustainability committees in achieving ESG targets. *Business Strategy and the Environment*, 33(3):2250–2268.

Adams, R. B., De Haan, J., Terjesen, S., and Van Ees, H. (2015). Board diversity: Moving the field forward. *Corporate Governance-An International Review*, 23(2):77–82.

Adams, R. B. and Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of financial economics*, 94(2):291–309.

Adams, R. B., Hermalin, B. E., and Weisbach, M. S. (2010). The role of boards of directors in corporate governance: A conceptual framework and survey. *Journal of economic literature*, 48(1):58–107.

Adamu, A. I., Yusuf, I., Bala, A. J., et al. (2024). Do Gender Diversity And Ceo Culture Affect Sustainability Performance In The Nigerian Banking Industry? *Journal of Central Banking Law and Institutions*, 3(1):37–56.

Aladwey, L. M. A. and Alsudays, R. A. (2023). Does the Cultural Dimension Influence the Relationship between Firm Value and Board Gender Diversity in Saudi Arabia, Mediated by ESG Scoring? *Journal of Risk and Financial Management*, 16(12):512.

AlJanadi, Y. (2025). Gender diversity and disclosure: a meta-analysis. *International Journal of Disclosure and Governance*, pages 1–15.

Alkhawaja, A., Hu, F., Johl, S., and Nadarajah, S. (2023). Board gender diversity, quotas, and ESG disclosure: Global evidence. *International Review of Financial Analysis*, 90:102823.

Alkurdi, A., Al Amosh, H., and Khatib, S. F. (2023). The mediating role of carbon emissions in the relationship between the board attributes and ESG performance: European evidence. *EuroMed Journal of Business*, (ahead-of-print).

Aly, D., Abdelqader, M., Darwish, T. K., Toporkiewicz, A., and Radwan, A. (2024). Board characteristics and sustainability in higher education institutions: The case of the United Kingdom. *Higher Education Quarterly*, page e12496.

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

Amorelli, M.-F. and García-Sánchez, I.-M. (2021). Trends in the dynamic evolution of board gender diversity and corporate social responsibility. *Corporate Social Responsibility and Environmental Management*, 28(2):537–554.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of mathematical statistics*, 20(1):46–63.

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

Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. (2010). On making causal claims: A review and recommendations. *The leadership quarterly*, 21(6):1086–1120.

Arayssi, M., Jizi, M., and Tabaja, H. H. (2020). The impact of board composition on the level of ESG disclosures in GCC countries. *Sustainability Accounting, Management and Policy Journal*, 11(1):137–161.

Atif, M., Hossain, M., Alam, M. S., and Goergen, M. (2021). Does board gender diversity affect renewable energy consumption? *Journal of Corporate Finance*, 66:101665.

Avramov, D., Cheng, S., Lioui, A., and Tarelli, A. (2022). Sustainable investing with ESG rating uncertainty. *Journal of financial economics*, 145(2):642–664.

Baldini, M., Maso, L. D., Liberatore, G., Mazzi, F., and Terzani, S. (2018). Role of country-and firm-level determinants in environmental, social, and governance disclosure. *Journal of Business Ethics*, 150:79–98.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful women: does exposure reduce bias? *The Quarterly Journal of Economics*, 124(4):1497–1540.

Bear, S., Rahman, N., and Post, C. (2010). The impact of board diversity and gender composition on corporate social responsibility and firm reputation. *Journal of Business Ethics*, 97:207–221.

Bel-Oms, I. and Segarra-Moliner, J. R. (2022). How do remuneration committees affect corporate social responsibility disclosure? Empirical evidence from an international perspective. *Sustainability*, 14(2):860.

Ben-Amar, W., Chang, M., and McIlkenny, P. (2017). Board gender diversity and corporate response to sustainability initiatives: Evidence from the carbon disclosure project. *Journal of business ethics*, 142(2):369–383.

Benaguid, O., Sbai, H., Meghouar, H., and Antari, O. (2023). Board gender diversity and CSR performance: A French study. *Cogent Business & Management*, 10(3):2247226.

Berg, F., Koelbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344.

Berrone, P., Fosfuri, A., and Gelabert, L. (2017). Does greenwashing pay off? Understanding the relationship between environmental actions and environmental legitimacy. *Journal of Business Ethics*, 144:363–379.

Bhatia, S. and Marwaha, D. (2022). The influence of board factors and gender diversity on the ESG disclosure score: a study on Indian companies. *Global Business Review*, 23(6):1544–1557.

Birindelli, G., Dell'Atti, S., Iannuzzi, A. P., and Savioli, M. (2018). Composition and activity of the board of directors: Impact on ESG performance in the banking system. *Sustainability*, 10(12):4699.

Blau, P. M. (1977). Inequality and Heterogeneity: A Primitive Theory of Social Structure. *American Journal of Sociology*, 83(1):26–54.

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

Borkowski, S. C. and Ugras, Y. J. (1998). Business students and ethics: A meta-analysis. *Journal of Business Ethics*, 17(11):1117–1127.

Boukattaya, S., Ftiti, Z., Jeriji, M., and Louhichi, W. (2024). Does committees' gender diversity affect firms' environmental and social performance? *Business Strategy and the Environment*, 33(4):3550–3568.

Boulouta, I. (2013). Hidden connections: The link between board gender diversity and corporate social performance. *Journal of Business Ethics*, 113(2):185–197.

Brammer, S., He, H., and Mellahi, K. (2015). Corporate social responsibility, employee organizational identification, and creative effort: The moderating impact of corporate ability. *Group & Organization Management*, 40(3):323–352.

Bruna, M. G., Đang, R., Ammari, A., and Houanti, L. (2021). The effect of board gender diversity on corporate social performance: An instrumental variable quantile regression approach. *Finance Research Letters*, 40:101734.

Cambrea, D. R., Paolone, F., and Cucari, N. (2023). Advisory or monitoring role in ESG scenario: Which women directors are more influential in the Italian context? *Business Strategy and the Environment*, 32(7):4299–4314.

Capizzi, V., Gioia, E., Giudici, G., and Tenca, F. (2021). The divergence of ESG ratings: An analysis of Italian listed companies. *Journal of Financial Management, Markets and Institutions*, 9(02):2150006.

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

Carrasco, A., Francoeur, C., Labelle, R., Laffarga, J., and Ruiz-Barbadillo, E. (2015). Appointing women to boards: is there a cultural bias? *Journal of Business Ethics*, 129:429–444.

Cazachevici, A., Havránek, T., and Horváth, R. (2020). Remittances and economic growth: A meta-analysis. *World Development*, 134:105021.

Chatterji, A. K., Durand, R., Levine, D. I., and Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic management journal*, 37(8):1597–1614.

Chebbi, K. and Ammer, M. A. (2022). Board composition and ESG disclosure in Saudi Arabia: The moderating role of corporate governance reforms. *Sustainability*, 14(19):12173.

Christensen, H. B., Hail, L., and Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of accounting studies*, 26(3):1176–1248.

Cohen, P. N. and Huffman, M. L. (2007). Working for the woman? Female managers and the gender wage gap. *American Sociological Review*, 72(5):681–704.

Csonka, A. and Milhomem, C. (2023). Women on boards and beyond. Progress report, MSCI. Available at: https://www.msci.com/documents/1296102/43943104/MSCI+Women+on+Boards+and+Beyond+2023+Progress+Report.pdf.

Cucari, N., Esposito De Falco, S., and Orlando, B. (2018). Diversity of board of directors and environmental social governance: Evidence from Italian listed companies. *Corporate social responsibility and environmental management*, 25(3):250–266.

Dabbebi, A., Lassoued, N., and Khanchel, I. (2022). Peering through the smokescreen: ESG disclosure and CEO personality. *Managerial and Decision Economics*, 43(7):3147–3164.

Dang, R., Hikkerova, L., Simioni, M., and Sahut, J. M. (2023a). How do women on corporate boards shape corporate social performance? Evidence drawn from semiparametric regression. *Annals of Operations Research*, 330(1):361–388.

Dang, R., Houanti, L., Sahut, J. M., and Simioni, M. (2021). Do women on corporate boards influence corporate social performance? A control function approach. *Finance Research Letters*, 39:101645.

Dang, R., Houanti, L., Simioni, M., and Sahut, J. M. (2023b). The role of endogeneity in the relationship between board gender diversity and corporate social performance: evidence from a control function method. *Annals of Operations Research*, pages 1–33.

De Andres, P., Azofra, V., and Lopez, F. (2005). Corporate boards in OECD countries: Size, composition, functioning and effectiveness. *Corporate Governance: An International Review*, 13(2):197–210.

De Masi, S., Słomka-Gołębiowska, A., Becagli, C., and Paci, A. (2021). Toward sustainable corporate behavior: The effect of the critical mass of female directors on environmental, social, and governance disclosure. *Business Strategy and the Environment*, 30(4):1865–1878.

Delmas, M. A. and Burbano, V. C. (2011). The drivers of greenwashing. *California management review*, 54(1):64–87.

Deloitte Insights (2023). Women in the Boardroom: A Global Perspective (8th Edition). Available at: https://www2.deloitte.com/sg/en/pages/cxo-programs/articles/women-in-the-boardroom-global-perspective-eighth-edition.html.

Dienes, D. and Velte, P. (2016). The impact of supervisory board composition on CSR reporting. Evidence from the German two-tier system. *Sustainability*, 8(1):63.

Diligent Institute (2022). A Few Good Women: Gender and Leadership in the Boardroom. Available at: https://www.diligentinstitute.com/report/gender-leadership-boardrooms-2022/.

Dorfleitner, G., Halbritter, G., and Nguyen, M. (2015). Measuring the level and risk of corporate responsibility–An empirical comparison of different ESG rating approaches. *Journal of Asset Management*, 16:450–466.

Doucouliagos, C. and Stanley, T. D. (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic surveys*, 27(2):316–339.

Doucouliagos, H. and Stanley, T. D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, 47(2):406–428.

Drucker, P. (2012). *The practice of management*. Routledge.

Eagly, A. H., Gartzia, L., and Carli, L. L. (2014). Female advantage: Revisited. In *The Oxford Handbook of Gender in Organizations*, volume 52. Oxford University Press.

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

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

Ellili, N. O. D. (2023). Impact of corporate governance on environmental, social, and governance disclosure: Any difference between financial and non-financial companies? *Corporate Social Responsibility and Environmental Management*, 30(2):858–873.

Equileap (2024). Gender Equality Report & Ranking – Emerging markets edition. Available at: https://equileap.com/wp-content/uploads/2024/04/Equileap_2024_Gender_Equality_Report_Emerging_Markets.pdf.

European Institure for Gender Equality (2023). Gender Equality Index 2023: Country Data. Available at: https://eige.europa.eu/modules/ custom/eige_gei/app/content/downloads/factsheets/EU_2023_factsheet.pdf.

European Investment Bank (2022). Support for Female Entrepreneurs: Survey Evidence for Why It Makes Sense. Available at: https://www.eib.org/en/publications/support-for-female-entrepreneurs-survey-evidence-for-why-it-makes-sense.

European Women on Boards (2024). Gender Diversity on European Boards. Available at: https://europeanwomenonboards.eu/wp-content/uploads/2024/03/EWOB-Toolkit-for-change-Gender-diversity-on-European-boards.pdf.

Fahad, P. and Rahman, P. M. (2020). Impact of corporate governance on CSR disclosure. *International Journal of Disclosure and Governance*, 17(2):155–167.

Fan, Y., Li, S., and Yang, W. (2024). The impact of the percentage of female directors on corporate ESG score. *Finance Research Letters*, 63:105376.

Farooque, O. A., Dahawy, K. M., Shehata, N. F., and Soliman, M. T. (2022). ESG disclosure, board diversity and ownership: did the revolution make a difference in Egypt? *Corporate Ownership & Control*, 19(2):67–80.

Fayyaz, U.-E.-R., Jalal, R. N.-U.-D., Venditti, M., and Minguez-Vera, A. (2023). Diverse boards and firm performance: The role of environmental, social and

governance disclosure. *Corporate Social Responsibility and Environmental Management*, 30(3):1457–1472.

Ferjančič, U., Ichev, R., Lončarski, I., Montariol, S., Pelicon, A., Pollak, S., Šuštar, K. S., Toman, A., Valentinčič, A., and Žnidaršič, M. (2024). Textual analysis of corporate sustainability reporting and corporate ESG scores. *International Review of Financial Analysis*, page 103669.

Fernández, C., Ley, E., and Steel, M. F. J. (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Econometrics*, 100(2):381–427.

Foley, S., Linnehan, F., Greenhaus, J. H., and Weer, C. H. (2006). The impact of gender similarity, racial similarity, and work culture on family-supportive supervision. *Group & Organization Management*, 31(4):420–441.

Forbes, D. P. and Milliken, F. J. (1999). Cognition and corporate governance: Understanding boards of directors as strategic decision-making groups. *Academy of management review*, 24(3):489–505.

Fragkiskos, F., Georgakopoulos, G., Lappas, T., and Stefanidis, A. (2023). Gender Diversity in the Boardroom and Corporate Performance: A Meta-Analysis. *British Journal of Management*, 34(3):1223–1246.

Franke, G. R., Crown, D. F., and Spake, D. F. (1997). Gender differences in ethical perceptions of business practices: A social role theory perspective. *Journal of Applied Psychology*, 82(6):920.

Fu, B., Wang, K., and Zhou, T. (2023). A Controversy in Sustainable Development: How Does Gender Diversity Affect the ESG Disclosure? In *International Conference on Economic Management and Green Development*, pages 669–678. Springer.

Furukawa, C. (2019). Publication bias under aggregation frictions: Theory, evidence, and a new correction method. Technical report, Kiel, Hamburg: ZBW–Leibniz Information Centre for Economics.

Gangi, F., Daniele, L. M., Varrone, N., Vicentini, F., and Coscia, M. (2021). Equity mutual funds' interest in the environmental, social and governance policies of target firms: Does gender diversity in management teams matter? *Corporate Social Responsibility and Environmental Management*, 28(3):1018–1031.

Gechert, S., Havránek, T., Havránková, Z., and Kolcunová, D. (2019). Death to the Cobb-Douglas production function. Technical report, IMK working paper.

Gechert, S., Havránek, T., Iršová, Z., and Kolcunová, D. (2022). Measuring capital-labor substitution: The importance of method choices and publication bias. *Review of Economic Dynamics*, 45:55–82.

George, E. I. (2010). Dilution priors: Compensating for model space redundancy. *IMS Collections: Borrowing Strength: Theory Powering Applications-A. Festschrift for Lawrence D. Brown*, 6(42):158–165.

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

Gerged, A. M., Tran, M., and Beddewela, E. S. (2023). Engendering pro-sustainable performance through a multi-layered gender diversity criterion: Evidence from the hospitality and tourism sector. *Journal of Travel Research*, 62(5):1047–1076.

Giannarakis, G. (2013). Determinants of corporate social responsibility disclosures: the case of the US companies. *International Journal of Information Systems and Change Management*, 6(3):205–221.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard educational review*, 47(4):481–517.

Gurol, B. and Lagasio, V. (2022). Women board members' impact on ESG disclosure with environment and social dimensions: evidence from the European banking sector. *Social Responsibility Journal*, 19(1):211–228.

Hambrick, D. C. and Mason, P. (1984). Academy of Management. *The Academy of Management Review*, 9:193.

Harjoto, M., Laksmana, I., and Lee, R. (2015). Board Diversity and Corporate Social Responsibility. *Journal of Business Ethics*, 132(4):641–660.

Harjoto, M. A. and Wang, Y. (2020). Board of directors network centrality and environmental, social and governance (ESG) performance. *Corporate Governance: The international journal of business in society*, 20(6):965–985.

Hasan, I., Horvath, R., and Mares, J. (2018). What type of finance matters for growth? Bayesian model averaging evidence. *The World Bank Economic Review*, 32(2):383–409.

Havránek, T., Iršová, Z., Janda, K., and Zilberman, D. (2015). Bank Capital and the Cost of Equity: A Meta-Analysis. *Journal of Financial Stability*, 19:55–74.

Havránek, T., Iršová, Z., Laslopová, L., and Zeynalova, O. (2021). Skilled and unskilled labor are less substitutable than commonly thought.

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

Havránek, T., Rusnak, M., and Sokolová, A. (2017). Habit formation in consumption: A meta-analysis. *European Economic Review*, 95:142–167.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255.

Heinemann, F., Moessinger, M.-D., and Yeter, M. (2018). Do fiscal rules constrain fiscal policy? A meta-regression-analysis. *European Journal of Political Economy*, 51:69–92.

Hillman, A. J., Shropshire, C., and Cannella Jr, A. A. (2007). Organizational predictors of women on corporate boards. *Academy of management journal*, 50(4):941–952.

Hoang, K. (2022). Silent but deadly: Political corruption and voluntary ESG disclosure in the United States. *Managerial and Decision Economics*, 43(7):2775–2793.

Hussain, N., Rigoni, U., and Orij, R. P. (2018). Corporate governance and sustainability performance: Analysis of triple bottom line performance. *Journal of business ethics*, 149:411–432.

Husted, B. W. and de Sousa-Filho, J. M. (2019). Board structure and environmental, social, and governance disclosure in Latin America. *Journal of Business Research*, 102:220–227.

Ibrahim, A. H. and Hanefah, M. M. (2016). Board diversity and corporate social responsibility in Jordan. *Journal of Financial Reporting and Accounting*, 14(2):279–298.

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

Iršová, Z., Doucouliagos, H., Havránek, T., and Stanley, T. (2023). Meta-analysis of social science research: A practitioner's guide. *Journal of Economic Surveys*.

Issa, A., Zaid, M. A., and Hanaysha, J. R. (2022). Exploring the relationship between female director's profile and sustainability performance: Evidence from the Middle East. *Managerial and Decision Economics*, 43(6):1980–2002.

Kamaludin, K., Ibrahim, I., Sundarasen, S., and Faizal, O. (2022). ESG in the boardroom: evidence from the Malaysian market. *International Journal of Corporate Social Responsibility*, 7(1):4.

Kanter, R. M. (2008). *Men and women of the corporation: New edition*. Basic books.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Khalid, F., Razzaq, A., Ming, J., and Razi, U. (2022). Firm characteristics, governance mechanisms, and ESG disclosure: how caring about sustainable concerns? *Environmental Science and Pollution Research*, 29(54):82064–82077.

Khatri, I. (2023). Board gender diversity and sustainability performance: Nordic evidence. *Corporate Social Responsibility and Environmental Management*, 30(3):1495–1507.

Khemakhem, H., Arroyo, P., and Montecinos, J. (2023). Gender diversity on board committees and ESG disclosure: evidence from Canada. *Journal of Management and Governance*, 27(4):1397–1422.

Konrad, A. M., Cannings, K., and Goldberg, C. B. (2010). Asymmetrical demography effects on psychological climate for gender diversity: Differential effects of leader gender and work unit gender composition among Swedish doctors. *Human Relations*, 63(11):1661–1685.

Konrad, A. M., Kramer, V., and Erkut, S. (2008). Critical mass:: The impact of three or more women on corporate boards. *Organizational dynamics*, 37(2):145–164.

Kravchenko, G., Brezovnik, B., Mlinaric, F., and Tselinko, I. (2023). Unlocking ESG Potential: The Interaction of Gender Diversity and Specialized Skills. *Unpublished*.

Kristianti, I. P. (2023). The Effect of Gender Diversity Boards on Sustainble Financial Reporting Practices. *JAK (Jurnal Akuntansi) Kajian Ilmiah Akuntansi*, 10(2):256–272.

Lavin, J. F. and Montecinos-Pearce, A. A. (2021). ESG disclosure in an emerging market: an empirical analysis of the influence of board characteristics and ownership structure. *Sustainability*, 13(19):10498.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4):651–674.

Li, C. and Wang, X. (2022). Local peer effects of corporate social responsibility. *Journal of Corporate Finance*, 73:102187.

Li, M., Huang, M., Wang, D., and Li, X. (2023). Star CEOs and ESG performance in China: An integrated view of role identity and role constraints logics. *Business Ethics, the Environment & Responsibility*, 32(4):1411–1428.

Liu, Y., Lei, L., and Buttner, E. H. (2020). Establishing the boundary conditions for female board directors' influence on firm performance through CSR. *Journal of Business Research*, 121:112–120.

Liu, Y., Zhang, F., and Zhang, H. (2024). CEO foreign experience and corporate environmental, social, and governance (ESG) performance. *Business Strategy and the Environment*, 33(4):3331–3355.

Lokuwaduge, C. S. D. S. and Heenetigala, K. (2017). Integrating environmental, social and governance (ESG) disclosure for a sustainable development: An Australian study. *Business Strategy and the Environment*, 26(4):438–450.

Lyu, W., Gao, Y., and Abdul Salam, Z. (2022). Application of BP Neural Network and Fixed Effect in the Relationship between CSR and Female Directors. *Journal of Mathematics*, 2022(1):7048186.

Mackenzie, C. (2007). Boards, incentives and corporate social responsibility: The case for a change of emphasis. *Corporate Governance: An International Review*, 15(5):935–943.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of econometrics*, 154(2):139–153.

Manita, R., Bruna, M. G., Dang, R., and Houanti, L. (2018). Board gender diversity and ESG disclosure: evidence from the USA. *Journal of Applied Accounting Research*, 19(2):206–224.

Martínez, M. d. C. V., Martín-Cervantes, P. A., and del Mar Miralles-Quirós, M. (2022). Sustainable development and the limits of gender policies on corporate boards in Europe. A comparative analysis between developed and emerging markets. *European Research on Management and Business Economics*, 28(1):100168.

Martiny, A., Taglialatela, J., Testa, F., and Iraldo, F. (2024). Determinants of environmental social and governance (ESG) performance: A systematic literature review. *Journal of Cleaner Production*, 456:142213.

McBrayer, G. A. (2018). Does persistence explain ESG disclosure decisions? *Corporate Social Responsibility and Environmental Management*, 25(6):1074–1086.

McGuinness, P. B., Vieito, J. P., and Wang, M. (2017). The role of board gender and foreign ownership in the CSR performance of Chinese listed firms. *Journal of Corporate Finance*, 42:75–99.

Minichilli, A., Zattoni, A., Nielsen, S., and Huse, M. (2012). Board task performance: An exploration of micro-and macro-level determinants of board effectiveness. *Journal of organizational behavior*, 33(2):193–215.

Miranda, B., Delgado, C., and Branco, M. C. (2023). Board characteristics, social trust and ESG performance in the European banking sector. *Journal of Risk and Financial Management*, 16(4):244.

Mooneeapen, O., Abhayawansa, S., and Mamode Khan, N. (2022). The influence of the country governance environment on corporate environmental, social and governance (ESG) performance. *Sustainability Accounting, Management and Policy Journal*, 13(4):953–985.

Nadeem, M., Zaman, R., and Saleem, I. (2017). Boardroom gender diversity and corporate sustainability practices: Evidence from Australian Securities Exchange listed firms. *Journal of Cleaner Production*, 149:874–885.

Naveed, K., Voinea, C. L., Ali, Z., Rauf, F., and Fratostiteanu, C. (2021). Board gender diversity and corporate social performance in different industry groups: Evidence from China. *Sustainability*, 13(6):3142.

Nekhili, M., Boukadhaba, A., and Nagati, H. (2021). The ESG–financial performance relationship: Does the type of employee board representation matter? *Corporate Governance: An International Review*, 29(2):134–161.

Nerantzidis, M., Tzeremes, P., Koutoupis, A., and Pourgias, A. (2022). Exploring the black box: Board gender diversity and corporate social performance. *Finance Research Letters*, 48:102987.

Nielsen, S. and Huse, M. (2010). Women directors' contribution to board decision-making and strategic involvement: The role of equality perception. *European Management Review*, 7(1):16–29.

Nuhu, Y. and Alam, A. (2024). Board characteristics and ESG disclosure in energy industry: evidence from emerging economies. *Journal of Financial Reporting and Accounting*, 22(1):7–28.

Özcan, İ. Ç. (2019). Determinants of environmental, social, and governance disclosure performance of publicly traded airports. *International journal of transport economics: Rivista internazionale di economia dei trasporti: XLVI, 3, 2019*, pages 77–92.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Paolone, F., Pozzoli, M., Chhabra, M., and Di Vaio, A. (2024). Cultural and gender diversity for ESG performance towards knowledge sharing: empirical evidence from European banks. *Journal of Knowledge Management*, 28(11):106–131.

Pavlova, E. and Gvetadze, S. (2023). Female access to finance: a survey of literature. EIF Research and Market Analysis Working Paper 2023/87, European Investment Fund.

Peng, H. and Chandarasupsang, T. (2023). The effect of female directors on ESG practice: Evidence from China. *International Journal of Financial Studies*, 11(2):66.

Pierli, G., Murmura, F., and Palazzi, F. (2022). Women and leadership: How do women leaders contribute to companies' sustainable choices? *Frontiers in Sustainability*, 3:930116.

Post, C., Rahman, N., and Rubow, E. (2011). Green governance: Boards of directors' composition and environmental corporate social responsibility. *Business & Society*, 50(1):189–223.

Puaschunder, J. M. (2019). The history of ethical, environmental, social, and governance-oriented investments as a key to sustainable prosperity in the finance world. *Public Integrity*, 21(2):161–181.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1995). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 90(430):773–795.

Rao, K. and Tilt, C. (2016). Board composition and corporate social responsibility: The role of diversity, gender, strategy and decision making. *Journal of business ethics*, 138:327–347.

Romano, M., Cirillo, A., Favino, C., and Netti, A. (2020). ESG (Environmental, Social and Governance) performance and board gender diversity: The moderating role of CEO duality. *Sustainability*, 12(21):9298.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., and Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, 19(1):4–60.

Schwartz, S. H. and Rubel, T. (2005). Sex differences in value priorities: cross-cultural and multimethod studies. *Journal of Personality and social Psychology*, 89(6):1010.

Sepulveda-Nuñez, M. D. D. C., Fong Reynoso, C., and Llamosas-Rosas, I. (2024). Board of directors effect on environmental, social and governance performance in publicly traded non-financial firms. *Corporate Governance: The International Journal of Business in Society*.

Setiani, E. P. and Novitasari, B. T. (2024). Exploring the Impact of Board Attributes on ESG Scores of Indonesian Companies. *Nominal Barometer Riset Akuntansi dan Manajemen*, 13(1):131–143.

Shahab, Y., Ntim, C. G., Chengang, Y., Ullah, F., and Fosu, S. (2018). Environmental policy, environmental performance, and financial distress in China: Do top management team characteristics matter? *Business Strategy and the Environment*, 27(8):1635–1652.

Shahbaz, M., Karaman, A. S., Kilic, M., and Uyar, A. (2020). Board attributes, CSR engagement, and corporate performance: what is the nexus in the energy sector? *Energy Policy*, 143:111582.

Shakil, M. H. (2021). Environmental, social and governance performance and financial risk: Moderating role of ESG controversies and board gender diversity. *Resources Policy*, 72:102144.

Shakil, M. H., Tasnia, M., and Mostafiz, M. I. (2021). Board gender diversity and environmental, social and governance performance of US banks: Moderating role of environmental, social and corporate governance controversies. *International Journal of Bank Marketing*, 39(4):661–677.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shin, T. (2012). The gender gap in executive compensation: The role of female directors and chief executive officers. *The Annals of the American Academy of Political and Social Science*, 639(1):258–278.

Singh, J., Sardana, V., and Singhania, S. (2024). Beyond compliance: the business case for gender diversity on boards and sustainability reporting in India. *International Journal of System Assurance Engineering and Management*, pages 1–10.

Słomka-Gołębiowska, A., De Masi, S., Zambelli, S., and Paci, A. (2023). Towards higher sustainability: If you want something done, ask a chairwoman. *Finance Research Letters*, 58:104308.

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review.

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

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

Stanley, T. D., Doucouliagos, H., and Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in medicine*, 36(10):1580–1598.

Stanley, T. D., Jarrell, S. B., and Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician*, 64(1):70–77.

Sterne, J. A. and Harbord, R. M. (2004). Funnel plots in meta-analysis. *The stata journal*, 4(2):127–141.

Sutiono, N. (2020). Board Gender Diversity, ESG, and Corporate Performance.

Torchia, M., Calabrò, A., and Huse, M. (2011). Women directors on corporate boards: From tokenism to critical mass. *Journal of business ethics*, 102:299–317.

Trireksani, T., Djajadikerta, H. G., Kamran, M., and Butt, P. (2024). The impact of country characteristics on board gender diversity and sustainability performance: a global perspective. *Sustainability*, 16(7):3057.

Uyar, A., Kilic, M., Koseoglu, M. A., Kuzey, C., and Karaman, A. S. (2020). The link among board characteristics, corporate social responsibility performance, and financial performance: Evidence from the hospitality and tourism industry. *Tourism Management Perspectives*, 35:100714.

Uyar, A., Kuzey, C., Kilic, M., and Karaman, A. S. (2021). Board structure, financial performance, corporate social responsibility performance, CSR committee, and CEO duality: Disentangling the connection in healthcare. *Corporate Social Responsibility and Environmental Management*, 28(6):1730–1748.

van Aert, R. C. and van Assen, M. A. (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 (oned on December 22, 2021).

Velte, P. (2016). Women on management board and ESG performance. *Journal of Global Responsibility*, 7(1):98–109.

Žigraiova, D. and Havránek, T. (2016). Bank competition and financial stability: Much ado about nothing? *Journal of Economic Surveys*, 30(5):944–981.

Wang, Y., Yekini, K., Babajide, B., and Kessy, M. (2022). Antecedents of corporate social responsibility disclosure: evidence from the UK extractive and retail sector. *International Journal of Accounting & Information Management*, 30(2):161–188.

Wasiuzzaman, S. and Wan Mohammad, W. M. (2020). Board gender diversity and transparency of environmental, social and governance disclosure: Evidence from Malaysia. *Managerial and decision economics*, 41(1):145–156.

Waterstraat, S., Kustner, C., and Koch, M. (2021). Does board composition taking account of sustainability expertise influence ESG ratings? An exploratory study of European banks. In *Society 5.0: First International Conference, Society 5.0 2021, Virtual Event, June 22–24, 2021, Revised Selected Papers 1*, pages 129–138. Springer.

Weber, O. (2014). Environmental, social and governance reporting in China. *Business Strategy and the Environment*, 23(5):303–317.

Williams, R. J. (2003). Women on corporate boards of directors and their influence on corporate philanthropy. *Journal of Business Ethics*, 42:1–10.

Wong, J. B. and Zhang, Q. (2022). Stock market reactions to adverse ESG disclosure via media channels. *The British Accounting Review*, 54(1):101045.

World Economic Forum (2020). Global gender gap report 2020. available at: https://www3.weforum.org/docs/wef_gggr_2020.pdf.

World Economic Forum (2024). Global gender gap report 2024. available at: https://www3.weforum.org/docs/wef_gggr_2024.pdf.

Wu, Q., Furuoka, F., and Lau, S. C. (2022). Corporate social responsibility and board gender diversity: a meta-analysis. *Management Research Review*, 45(7):956–983.

Xiaomei, Z., Sallehuddin, M. R., Saad, R. M., and Lu, Z. (2021). Gender diversity, political connections and environmental, social and governance performance in China: The moderating role of tenure. *Studies of Applied Economics*, 39(10).

Yan, J., Huang, Y., and Liao, X. (2024). Help or hindrance? The impact of female executives on corporate ESG performance in China. *Journal of Cleaner Production*, 437:140614.

Yarram, S. R. and Adapa, S. (2021). Board gender diversity and corporate social responsibility: Is there a case for critical mass? *Journal of cleaner production*, 278:123319.

Yilmaz, M. K., Hacioglu, U., Nantembelele, F. A., and Sowe, S. (2021). Corporate board diversity and its impact on the social performance of companies from emerging economies. *Global Business and Organizational Excellence*, 41(1):6–20.

Yorke, S. M., Donkor, A., and Appiagyei, K. (2023). Experts on boards audit committee and sustainability performance: The role of gender. *Journal of cleaner production*, 414:137553.

Zeugner, S. (2011). Bayesian model averaging with BMS. *Tutorial to the R-package BMS 1e30*.

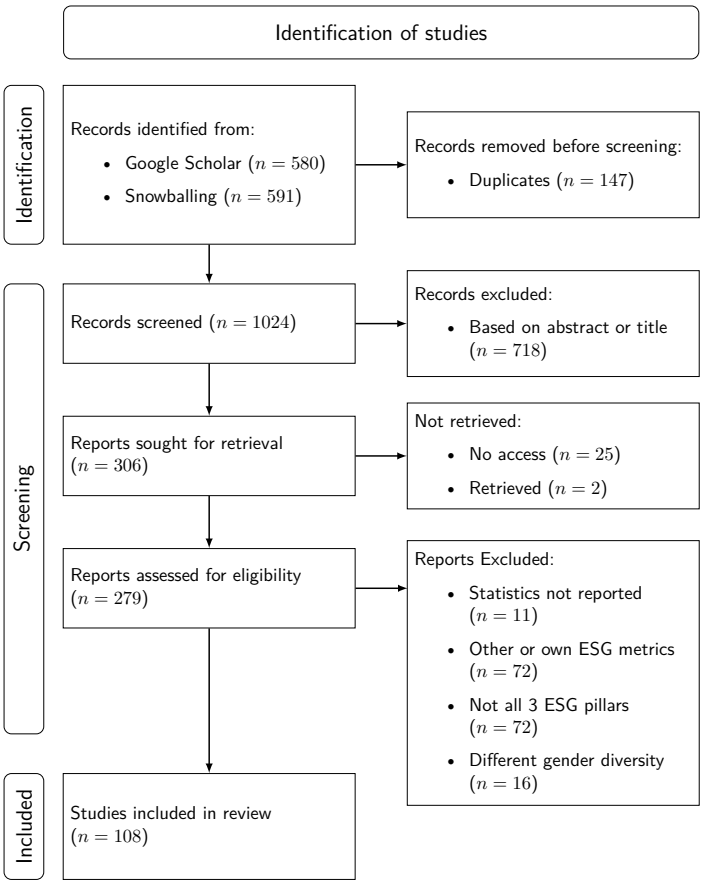
Zhang, D. (2022). Do heterogenous subsides work differently on environmental innovation? A mechanism exploration approach. *Energy Economics*, 114:106233.

Zhang, D. (2023). Subsidy expiration and greenwashing decision: Is there a role of bankruptcy risk? *Energy Economics*, 118:106530.

Zimmermann, C. (2013). Academic rankings with RePEc. *Econometrics*, 1(3):249–280.

Appendix A
Additional Information for Data Collection

Figure A.1: PRISMA flow diagram

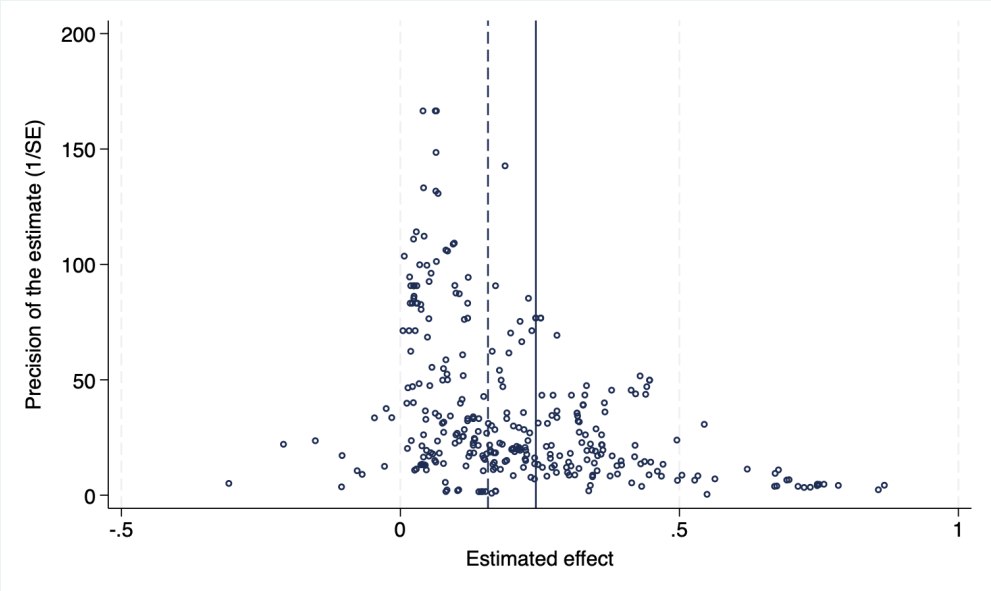


Notes: The figure shows a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram constructed in line with the identification of studies. The diagram was created based on the template by Page et al. (2021).

Appendix B
Robustness Checks for Publication Bias

Tests of publication bias on subsample of preferred estimates

Figure B.1: Funnel plot for a subsample of preferred estimates



Notes: In the absence of publication bias, the most precise estimates are expected to cluster around the mean effect. Less precise estimates should be symmetrically distributed around it. The figure suggests an asymmetry with the most precise estimates reported between zero and mean effect, denoted by the solid vertical line. The dashed line represents the median effect. Extreme outliers are omitted from the figure but remain included in all statistical tests.

Table B.1: Linear tests for a subsample of of preferred estimates

	OLS	Between Effects	Study Weight	Precision Weight	IV
Publication Bias	1.044***	0.911***	1.128**	2.474***	2.394***
(*Standard Error*)	(0.345)	(0.150)	(0.505)	(0.486)	(0.621)
	[0.292, 2.129]		[â'0.134, 2.29]	[1.189, 3.628]	{0.942, 4.830}
Mean Beyond Bias	0.150***	0.151***	0.188***	0.0369*	0.043
(*Constant*)	(0.025)	(0.032)	(0.043)	(0.022)	(0.043)
	[0.098, 0.202]		[0.101, 0.285]	[-0.004, 0.122]	{-0.042, 0.125}
First-stage robust F-stat					11.83
Observations	454	454	454	454	454
Studies	102	102	102	102	102

Notes: The results of the regression $E_{is} = \beta_0 + \beta_1 \cdot \text{SE}(E_{is}) + \varepsilon_{is}$, where E_{is} denotes the i-th estimated effect in the study s, and $\text{SE}(E_{is})$ denotes its standard error. 95% confidence intervals from wild bootstrap (Roodman et al., 2019) are shown in square brackets. For IV, the Anderson-Rubin 95% confidence interval recommended by Anderson and Rubin (1949) is shown in curly brackets. Standard errors and confidence intervals are enclosed in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Non-linear tests for a subsample of preferred estimates

	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		0.014		8.441*	
		(0.012)		(5.057)	
Effect Beyond Bias	0.024***	0.089	0.062***	0.023***	0.056***
	(0.006)	(0.087)	(0.007)	(0.003)	(0.019)
# of estimates			103		
% of information			31,7%		
Observations	454	454	454	454	454

Notes: WAAP denotes weighted average of adequately powered estimates (Ioannidis et al. (2017)); Selection model denotes technique due Andrews and Kasy (2019); Stem denotes the stem-based technique (Furukawa (2019)); Kink denotes the endogenous kink model (Bom and Rachinger (2019)); p-uniform* denotes the technique due to van Aert and van Assen (2021). Standard errors and confidence intervals are enclosed in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

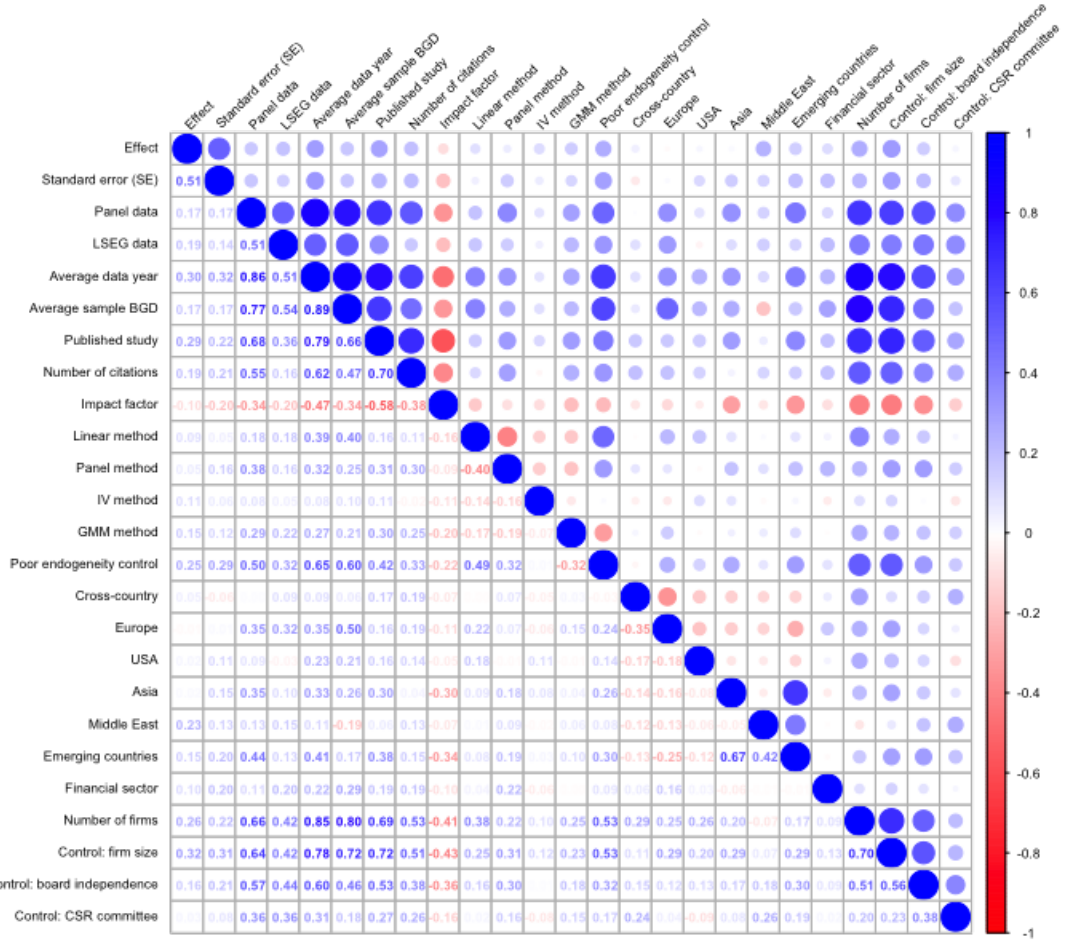
Table B.3: Caliper test for a subsample of preferred estimates

Critical threshold of t-statistic	0	1.96
Caliper size: 35%	0.941***	0.657***
Share of estimates above the threshold	(0.059)	(0.081)
	N=17	N=35
Caliper size: 45%	0.85***	0.66***
Share of estimates above the threshold	(0.082)	(0.067)
	N=20	N=50

Notes: The reported shares are statistically significantly above the 0.5 threshold. Standard errors in parentheses. The significance levels are indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

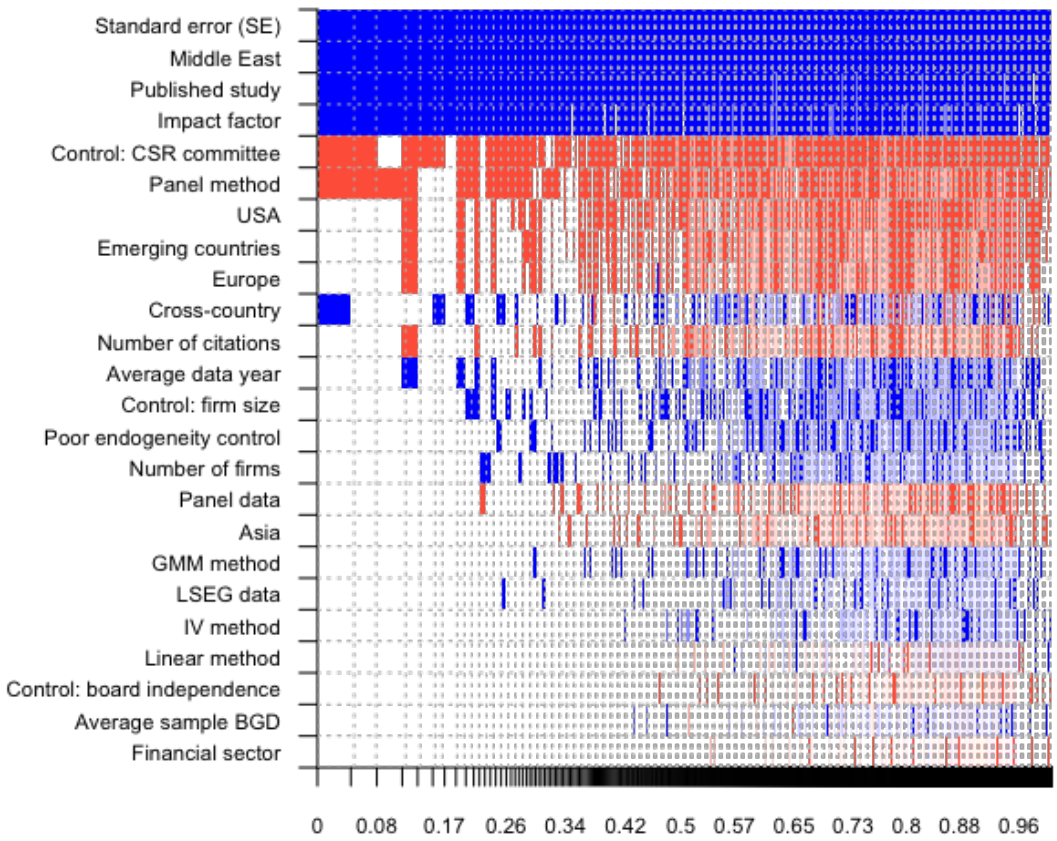
Appendix C
BMA Robustness Checks and Diagnostics

Figure C.1: Correlations between potential predictors of heterogeneity



Notes: The figure displays correlation coefficients for variables collected to study heterogeneity in estimated effect.

Figure C.2: Model inclusion in robustness BMA



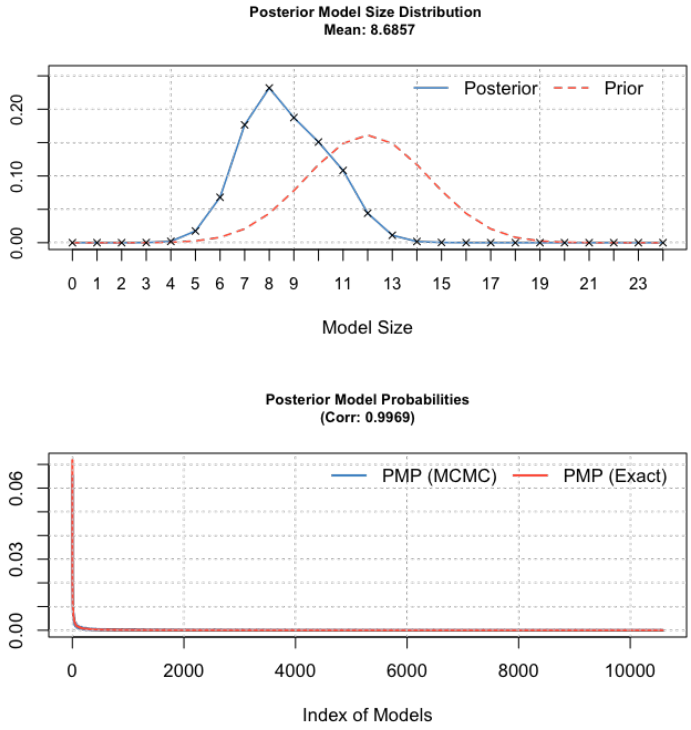
Notes: The figure shows results of the robustness BMA estimation reported in Table C.1 based on best 14421 models (g-prior = BRIC, model prior = random). The vertical axis ranks the explanatory variables based on their posterior inclusion probabilities, from highest at the top to lowest at the bottom. The horizontal axis represents the cumulative posterior model probability. A blue (darker) color indicates positive effect, red (lighter) color negative effect and absence of color denotes non-inclusion of the corresponding variable in the respective model. Variable descriptions in Table 6.1 and further technical diagnostic information in Appendix C Table C.3 and Figure C.4.

Table C.1: Results of robustness BMA and FMA

	Bayesian Model Averaging (BRIC and random priors)			**Frequentist Model Averaging**		
	P. Mean	*P. SD*	*PIP*	*Coef.*	*SD*	*p-value*
Standard error (SE)	0.887	0.078	1.000	0.801	0.081	0.000
Data characteristics						
Panel data	-0.018	0.050	0.152	-0.073	0.067	0.277
LSEG data	0.003	0.014	0.069	0.025	0.033	0.449
Average data year	0.009	0.017	0.250	0.061	0.024	0.011
Average sample BGD	0.001	0.008	0.042	-0.040	0.037	0.282
Publication characteristics						
Published	0.237	0.072	0.973	0.164	0.057	0.004
Citations	-0.005	0.010	0.252	-0.021	0.009	0.023
Impact factor	0.024	0.010	0.913	-0.003	-0.001	0.078
Estimation techniques						
Linear estimation	-0.001	0.014	0.045	-0.024	0.067	0.718
Panel method	-0.062	0.050	0.682	-0.073	0.067	0.273
Instrumental variable	0.006	0.029	0.061	0.018	0.085	0.831
GMM	0.013	0.047	0.099	0.078	0.082	0.340
Design of the analysis						
Poor endog. control	0.018	0.046	0.176	0.086	0.050	0.084
Global	0.012	0.067	0.294	-0.191	0.068	0.005
Europe	-0.057	0.091	0.358	-0.322	0.068	0.000
USA	-0.090	0.117	0.457	-0.379	0.074	0.000
Asia	-0.016	0.047	0.139	-0.181	0.064	0.005
Middle East	0.379	0.086	0.998	0.281	0.100	0.005
Emerging markets	-0.075	0.104	0.410	-0.273	0.070	0.000
Financial sector	-0.001	0.008	0.023	0.004	0.054	0.940
Number of firms	0.003	0.008	0.154	0.001	0.012	0.920
Control variables						
Firm size	0.025	0.050	0.238	0.1349	0.0483	0.005
Board independence	-0.002	0.011	0.043	-0.0465	0.0331	0.160
CSR committee	-0.084	0.051	0.800	-0.0881	0.0327	0.007

Notes: The table displays results for robustness BMA robustness model and frequentist model averaging. Robustness BMA uses the BRIC g-prior proposed by Fernández et al. (2001) and the random model prior due to Ley and Steel (2009). Variables are described in Table 6.1. P. Mean = Posterior Mean. P. SD = Posterior Standard Deviation. PIP = Posterior Inclusion Probability.

Figure C.3: Model size and convergence for the baseline BMA



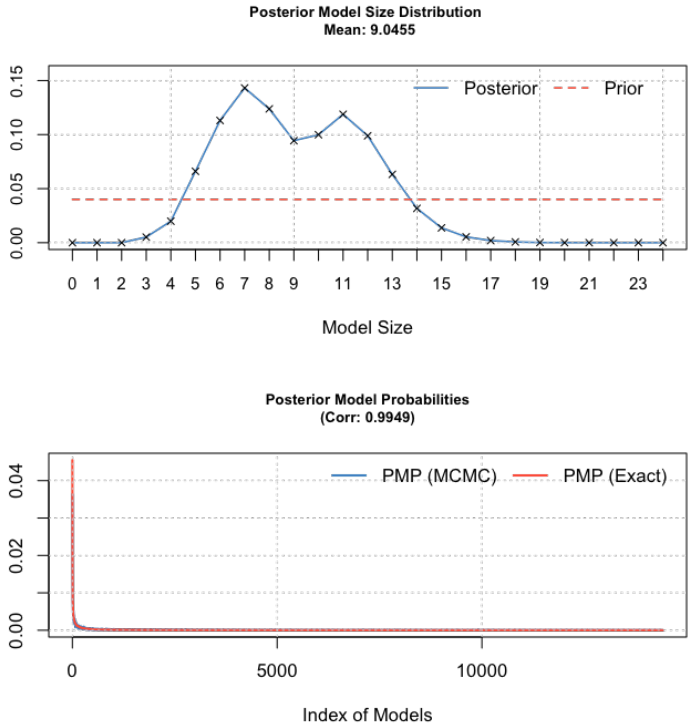
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the baseline BMA estimation reported in Table 6.2 using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.

Table C.2: Diagnostics of the baseline BMA (UIP and dilution priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.6857	$3 \cdot 10^5$	$1 \cdot 10^5$	2.18 mins	76,355
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	0.46%	100%	0.9969	547
Model prior	*g-prior*	*Shrinkage-stats*		
Dilution / 12	UIP	Av = 0.9982		

Notes: The table shows technical diagnostic information of baseline BMA estimation reported in Table 6.2 using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.

Figure C.4: Model size and convergence for the robustness BMA



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the robustness BMA estimation reported in Table C.1 using the BRIC g-prior proposed by Fernández et al. (2001) and the random model prior due to Ley and Steel (2009).

Table C.3: Diagnostics of the alternative BMA (BRIC and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
9.0455	$3 \cdot 10^5$	$1 \cdot 10^5$	1.26 mins	82,032
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$1.7 \cdot 10^7$	0.49%	100%	0.9949	547
Model prior	*g-prior*	*Shrinkage-stats*		
Random / 12	BRIC	Av = 0.9983		

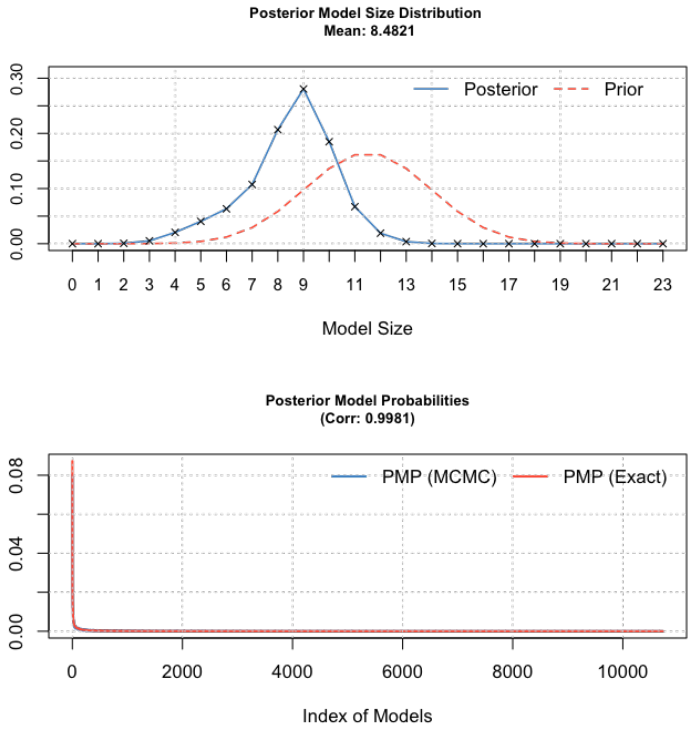
Notes: The table shows technical diagnostic information of robustness BMA estimation reported in Table C.1 using the BRIC g-prior proposed by Fernández et al. (2001) and the random model prior due to Ley and Steel (2009).

Table C.4: Results of subsample robustness BMA and frequentist check estimations

	Bayesian Model Averaging (baseline model)			**Stepwise regression (frequentist check)**		
	P. Mean	*P. SD*	*PIP*	*Coef.*	*SE*	*p-value*
Standard error (SE)	0.889	0.085	1.000	0.905	0.070	0.000
Data characteristics						
Panel data	-0.004	0.025	0.042			
LSEG	0.006	0.020	0.118			
Average data year	0.008	0.018	0.193			
Average sample BGD	-0.000	0.005	0.017			
Publication characteristics						
Published	0.147	0.100	0.754	0.153	0.042	0.000
Citations	-0.001	0.004	0.068			
Impact factor	0.009	0.011	0.467			
Estimation techniques						
Linear estimation	0.003	0.015	0.063			
Panel method	-0.051	0.047	0.608	-0.080	0.029	0.005
Instrumental variable	0.012	0.041	0.112			
GMM	0.021	0.058	0.152			
Design of the analysis						
Poor endog. control	0.018	0.046	0.176			
Global	-0.034	0.089	0.191			
Europe	-0.132	0.109	0.836	-0.113	0.031	0.000
USA	-0.178	0.116	0.882	-0.173	0.044	0.000
Asia	-0.178	0.078	0.920	-0.192	0.043	0.000
Emerging markets	-0.023	0.070	0.122			
Financial sector	-0.001	0.010	0.034			
Number of firms	0.000	0.003	0.033			
Control variables						
Firm size	0.123	0.080	0.767	0.163	0.042	0.000
Board independence	-0.005	0.019	0.092			
CSR committee	-0.072	0.047	0.773	-0.095	0.029	0.001
Studies	102			102		
Observations	517			517		

Notes: The table displays results for subsample robustness BMA model and frequentist check that includes only variables with PIP > 0.5 in the BMA model. BMA uses the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively. Variables are described in Table 6.1. P. Mean = Posterior Mean. P. SD = Posterior Standard Deviation. PIP = Posterior Inclusion Probability.

Figure C.5: Model size and convergence for the subsample robustness BMA



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the robustness BMA estimation reported in Table C.4 using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.

Table C.5: Diagnostics of the subsample robustness BMA

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.4821	$3 \cdot 10^5$	$1 \cdot 10^5$	2.13 mins	76,902
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
2^{23}	0.92%	100%	0.9981	545
Model prior	*g-prior*	*Shrinkage-stats*		
Dilution / 11.5	UIP	Av = 0.9982		

Notes: The table shows technical diagnostic information of subsample robustness BMA estimation reported in Table C.4 using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.