INSTITUTE
OF ECONOMIC STUDIES
Faculty of Social Sciences
Charles University

FEMALE LEADERSHIP AND FINANCIAL PERFORMANCE: A META-ANALYSIS

Katarina Gomoryova

IES Working Paper 6/2024

Institute of Economic Studies,
Faculty of Social Sciences,
Charles University in Prague

[UK FSV – IES]

Opletalova 26
CZ-110 00, Prague
E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Institut ekonomických studií
Fakulta sociálních věd
Univerzita Karlova v Praze

Opletalova 26
110 00 Praha 1

E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Bibliographic information:
Gomoryova K. (2024): " Female Leadership and Financial Performance: A Meta-Analysis" IES Working Papers 6/2024. IES FSV. Charles University.

Female Leadership and Financial Performance: A Meta-Analysis

Katarina Gomoryova

Charles University, Prague, Czech Republic
E-mail: kgomoryova30@gmail.com

January 2024

Abstract:
Is female leadership the secret ingredient to financial prosperity? This question has been the subject of extensive research, yet the findings remain inconclusive. We aim to provide a comprehensive understanding of this relationship employing contemporary techniques on the up-to-date dataset comprising 1,131 estimates gathered from 96 distinct studies. We address the pervasive issue of publication bias resulting in the mild preference for positive outcomes. After filtering out this bias, the study finds a negligible mean effect estimate, suggesting that the impact of women in leadership on financial performance is minimal. We further explore the potential factors that could account for variations in the estimated effects across different studies. Utilising Bayesian Model Averaging, weighted by the inverse number of estimates, we identify thirteen significant moderators that influence the relationship under study. Among these, the proportion of female authors, the impact factor of the journal, the duality of the CEO role, and the tenure of leaders are found to exert the most positive influence on the effect. Conversely, the age of leaders pushes effect the most in the opposite direction. Other influential factors include the publication status of the article, the number of variables used in the study, publication bias, the use of random estimation and matching approaches, the use of accounting-based financial measures, focus on the emerging market, and the representation of the leadership variable as a proportion.

JEL: J23, J24, J31
Keywords: meta-analysis, publication bias, Bayesian Model Averaging, female leadership, gender diversity, financial performance

Acknowledgement: Gomoryova acknowledges support from The Charles University Grant Agency (grant #217623) from which this project has received funding. This output is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 870245.

1 Introduction

The role of women in leadership positions and its impact on financial performance has been a topic of significant interest in recent years. While the proportion of women in such positions is on the rise, they remain globally underrepresented with women constituting 15% of all CEOs and managing directors in 2019 and 26% in 2021. This demographic shift and its potential implications for organisational performance have fuelled the focus on this issue.

Women bring a unique skill set to the table, enhance supervisory and monitoring functions (Bennouri et al., 2018) and contribute to risk mitigation within firms (Steffensmeier et al., 2013). Yet, even with the growing body of research and the business case advocating for the unique value added of women, a consensus on the actual effect remains inconclusive.

For instance, Carter et al. (2003) discover a positive association between the number of female board members and the value of the firm. Similarly, Dezsö and Ross (2012) find that companies with women in top management positions demonstrate superior financial performance. Adams and Ferreira (2009) highlight how women's skill sets influence the effectiveness of companies. However, they also indicate that the overall effect of gender diversity on financial performance is negative. Whereas Gregory-Smith et al. (2014) find no evidence to support the claim that an increase in the percentage of women on boards boosts financial performance. Despite ongoing debate, if women perform at least at par with men, their representation in the upper echelons of organisations worldwide should be more balanced.

This study aims to provide a comprehensive understanding of this relationship through meta-analytic methods. Meta-analysis is a powerful tool that allows for the synthesis of findings from multiple studies, providing more robust and reliable results. We recognise four core meta-analyses trying to grasp this complex relationship: Eagly and Carli (2003), Pletzer et al. (2015), Post and Byron (2015) and Hoobler et al. (2018). Even though they provide valuable contributions, they do not thoroughly account for publication bias or investigate heterogeneity behind estimates.

Our analysis incorporates 1,131 estimates from 96 distinct studies, along with their standard errors and variables that underscore the differences between the studies. Primary studies employ distinct financial ratios, often calculated without a shared basis or without disclosing their calculation methods. Consequently, we opt to utilise partial correlation coefficients in our work.

Publication bias is a pervasive issue that presents a substantial challenge in most published literature (Stanley, 2005). The choice to publish a study often depends on the statistical significance of its findings, which can prompt authors to adjust sample sizes and model specifications to reach this significance (Gerber and Malhotra, 2008). To ascertain the presence of publication bias in the gathered estimates, we employ several contemporary statistical tests. We start our examination with visual instrument known as Funnel Plot. Then we execute the FAT-PAT with following specifications: OLS, between study variance, weighted by inverse number of estimates and by precision. Subsequently, we utilise series of non-linear techniques such as the Weighted Average of Adequately Powered by Ioannidis et al. (2017), the Selection model as per Andrews and Kasy (2019), the Stem-based method as proposed by Furukawa (2019) and Endogenous kink model by Bom and Rachinger (2019). Lastly, we adopt methods allowing for endogeneity, such as an instrumental variable approach, and Caliper tests (Gerber and Malhotra, 2008). The outcomes of these examinations suggest only a slight indication of positive publication bias. Once we eliminate this bias, we arrive at an negligible mean effect estimate. This observation is further corroborated by our best practice estimate, an estimate derived from a synthetic study with pre-established optimal conditions, which results in higher value but still close to zero.

It is plausible that the estimated effects of primary studies vary due to factors beyond publication bias. These factors could encompass the unique settings of the studies, methodology, and a host of other elements. As the baseline model we perform Bayesian Model Averaging weighted by inverse number of estimates. In total, we control for 37 variables, and the baseline model distinguishes thirteen as significant moderators of the relationship under study. This analytical exercise softens the positive publication bias when we control for additional study characteristics. Factors that positively influence the relationship include the female ratio among authors, the impact factor of the journal, whether the study was published in a peer-reviewed journal, the use of a random effects model for estimation, the proportion of female leadership, the tenure of leaders, and the duality of the CEO role. Conversely, factors such as the number of variables in the primary analysis, analysis performed on matched datasets, the use of accounting-based financial measures, or the age of leaders tend to decrease the effect.

Understanding the relationship between female leadership and financial performance is not only critical for academic purposes but also has practical implications. It informs organisational

policies and practices regarding gender diversity and inclusion, leadership development, and succession planning. This meta-analysis aims to contribute to this understanding, providing insights that can guide future research and practice in this area.

2 The Dataset

For a comprehensive meta-analysis, it is essential to compile a dataset containing diverse variables from studies related to the primary subject. Our study focuses on research examining gender diversity and its impact on an organisation's financial performance. While gender diversity has been extensively explored across disciplines such as behavioural economics, psychology, and sociology, reviewing all pertinent literature would be highly time-consuming. Consequently, this analysis primarily concentrates on studies that specifically investigate women in high-ranking positions, such as board members, top-level management, or CEOs, rather than considering gender diversity in the broader context of the entire organisation or industry.

2.1 Compilation of Data

In accordance with the guidelines proposed by Stanley and Doucouliagos (2012), the selection process for studies to be included in the meta-analysis begins with an online search using Google Scholar. This platform serves as an extensive search engine with access to the full text of a vast number of studies. The literature search was limited to studies written in English to ensure a complete understanding of the reported findings. Given that the topic under investigation is not confined to a specific country and most academic papers in the studied field are published in English, excluding non-English literature does not significantly influence the results. Moreover, only studies accessible through a standard university licence are included in the final selection. In line with Stanley (2001), no study is excluded based on its publication status. Our final sample comprises research articles, working papers, theses, and doctoral dissertations.

The data search was conducted in February 2023, using the phrase *"female leadership and financial performance"*. The query generated 2,350,000 results, of which the first 500 were examined. We then refined the search to studies published in the last four years to encompass the most recent research. Further, we employed a snowballing technique, incorporating papers based on the references. This process identified 152 studies as potential candidates for the

meta-analysis. Each study had to fulfil the following selection criteria to qualify for inclusion in the meta-analysis: employ quantitative methods, utilise a financial outcome as the dependent variable, provide a clear coefficient of gender diversity, present standard errors of the coefficients or furnish other statistics from which standard errors are obtained, i.e. t-statistics and p-values, and report the number of observations.

On top of that, each study from the latest meta-analysis by Hoobler et al. (2018) that satisfied the criteria and was not already part of the sample was added to the list of final studies.

We faced several challenges during the data collection process, including incomplete documentation in the original research papers. These issues ranged from a lack of explicit details about uncertainty measures to errors in decimal point placement. Furthermore, inadequate explanations of the control variables or analytical methods used in the original regression heightened the risk of misinterpretation. The data compilation phase was completed in April 2023, and the resulting dataset is available upon request. Overall, the dataset encompasses 96 studies and includes a total of 1,131 estimates. The list of included studies is shown in Table A1.

2.2 Tranformation of Data

We move forward with necessary data adjustments upon completing the data collection process. This step ensures that the dataset includes comparable effect estimates. Certain studies scrutinise the non-linear association between female leadership and financial performance via a quadratic term. To address the issue of two estimates representing the same effect, we adopt the approach from Zigraiova and Havranek (2016) and linearise the investigated effect as follows:

$$\hat{\beta}_{is} = \hat{\beta}_{lis} + 2\hat{\beta}_{qis}\overline{x}_{es}, \tag{1}$$

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

Where β_{lis} denotes the coefficient for the linear estimate of female leadership, β_{qis} represents the estimate for the quadratic term of female leadership, β_{es} indicates the sample mean of female leadership in a given study and $\overline{x}_{es}$ represents the sample mean of the interacted variable.

Additionally, SE(β_{lis}) refers to the standard error for the linear term estimate, while SE(β_{qis}) refers to the standard error for the quadratic term estimate.

Furthermore, several studies in our dataset estimate an interaction term between female leadership and other variables, such as independence of boards (Kweh et al., 2019), CEO duality (Terjesen et al., 2016), and team size (Abdelzaher and Abdelzaher, 2019). In line with Cazachevici et al. (2020), we compute the average marginal effect of female leaders on the performance of the company and the associated standard errors by employing the delta method:

$$ME_{is} = \hat{\beta}_{lis} + \hat{\beta}_{tis}\overline{x}_{is}, \tag{3}$$

$$SE(ME_{is}) = \sqrt{SE(\hat{\beta}_{lis})^2 + SE(\hat{\beta}_{tis})^2\overline{x}_{is}}. \tag{4}$$

In this context, ME_{is} denotes the computed marginal effects of female leadership, β_{lis} signifies the estimated linear effect size of females in upper echelons, β_{tis} represents the coefficient of interaction term, and $\overline{x}_{is}$ is the mean value of the interacted variable. While SE(β_{lis}) and SE(β_{tis}) are standard errors associated with the variables of interest and the interacted variable. respectively. It is worth noting that certain studies lack summary statistics or uncertainty measures for quadratic or interaction terms, which makes the transformation and subsequent estimation impossible.

Two primary studies adopt a different approach in utilising dummy variables for the gender of the CEO compared to the remaining studies. Instead of using a dummy variable that takes the value of one for females and zero for males, they employ the opposite variation. To ensure consistent interpretation of the collected estimates regarding the impact of women in top positions on a company's financial performance, we adjust the coefficients by reversing their signs to make them comparable to the rest.

Despite the modifications detailed above, the collected estimates still exhibit differences in their econometric specifications and units of measurement. Adhering to the standardisation method adopted by Zigraiova and Havranek (2016) and Doucouliagos and Laroche (2003), we normalise the effect sizes of the gathered estimates by transforming them into partial correlation coefficients (PCCs). PCCs render the reported research findings directly comparable as they offer a measure without units that indicates the intensity and direction of the correlation, while holding other variables constant. They can fall within the interval [-1,1]. The signs of

PCCs correspond to the modified signs of the estimates. We determine the partial correlation coefficient and the standard error using the following formulas:

$$PCC_{is} = \frac{t_{is}}{\sqrt{(t^2)_{is} + df_{is}}}, \tag{5}$$

$$SE(PCC)_{is} = \sqrt{\frac{1 - (PCC^2)_{is}}{df_{is}}}. \tag{6}$$

Where, PCC_{is} signifies the partial correlation coefficient, $SE(PCC)_{is}$ is its associated standard error, t_{is} represents the t-statistic, and df_{is} indicates degrees of freedom.

The final stage of data transformation comprises a few modifications. Studies by Liu et al. (2014) and Xie et al. (2020) present zero standard errors for specific estimates, precluding the calculation of PCCs. To address this issue, we replace standard errors equal to 0 with 0.001, an adequately low value.

In cases where only asterisks are provided to report the significance of coefficients, we make the following adjustments: for studies indicating 5% significance, we approximate the t-statistics value as 2.27, while for those reporting 10% significance, we approximate the value as 1.8. We do not report t-statistics for coefficients where asterisks refer to 1% significance.

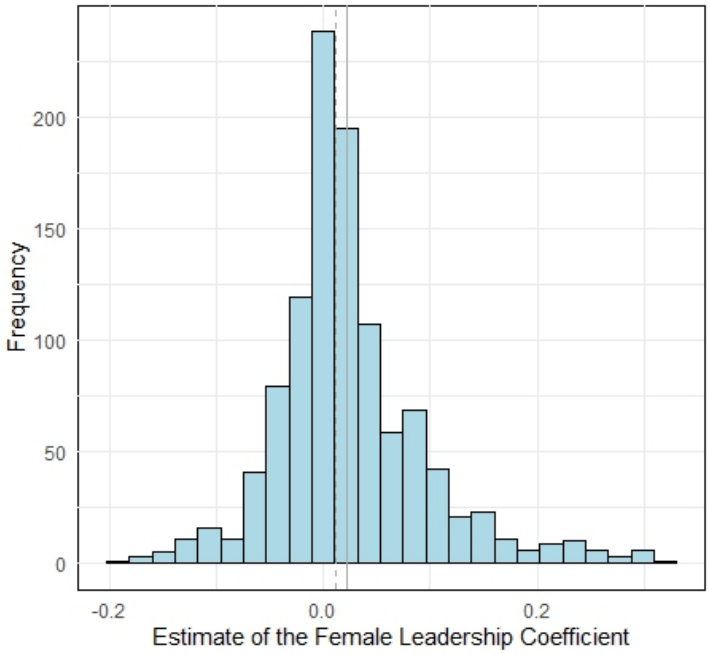
In conclusion, despite our rigorous efforts to clean the data, some extreme values of calculated PCCs, along with their associated standard errors and t-statistics, remain. In order to preserve the insights these observations provide without distorting the outcomes, we substitute these outliers and their measures of uncertainty with values that have been winsorised at the 1% level.

2.3 Descriptive Evidence

Before delving into the analysis of publication bias and heterogeneity, we present summary statistics to provide a comprehensive overview of the dataset. After carefully selecting the relevant studies and making necessary adjustments to the data, our dataset consists of 1,131 partial correlation coefficients extracted from 96 primary studies examining the impact of female leaders on the financial performance of companies. Our sample spans from 1997 to 2023, encompassing studies with the most recent publication dates. As depicted in Figure A1, there

is a discernible rise in the number of studies post-2010, indicating an increased interest in this research field in the last decade.

Figure 1: Distribution of calculated PCCs



Notes: The figure illustrates distribution of calculated PCCs. The mean is expressed by solid line and median by dashed line.

Figure 1 illustrates the distribution of PCC estimates of the examined relationship. As can be seen in Table 1 PCC values range from -0.2159 to 0.3358, with a mean of 0.0224 and a median of 0.012. The fact that the mean is higher than the median suggests a right-skewed data distribution. However, relying solely on a simple average would not provide a thorough understanding, as it would heavily favour primary studies with a higher count of effect estimates. In our dataset, we note variations in the number of estimates reported by different studies. For example, while eight studies report only one estimate, others like Flabbi et al. (2017) and Liu et al. (2014) report 95 and 92 estimates, respectively. To counter this imbalance, we use a weighting approach, giving more weight to studies with fewer reported estimates. Using the inverse of the number of estimates reported by each study, we obtain a weighted mean of PCC equal to 0.0335. This approach yields a more equitable representation of the overall effect across studies; however, according to Doucouliagos (2011), either means is considered a small effect.

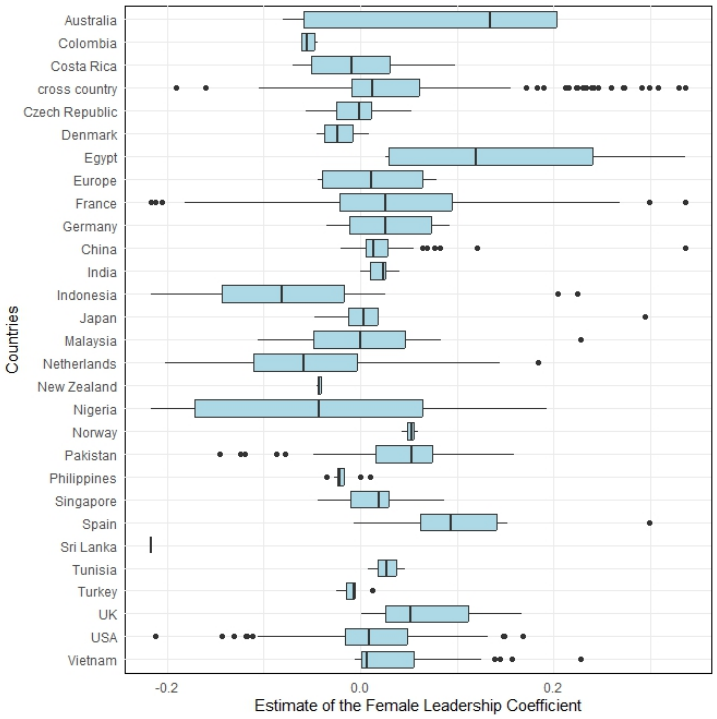
Figure 2 displays box plots illustrating a moderate variation of estimates across different countries. The boxes represent the interquartile range, encompassing from the 25th to the 75th

Table 1: Descriptive statistics for partial correlation coefficients

	Unweighted mean	Weighted mean	Median	SD	Minimum	Maximum
Partial Correlation Coefficient	0.0224	0.0335	0.012	0.0843	-0.2159	0.3358
Observations	1,131	1,131	1,131	1,131	1,131	1,131

percentile, while the whiskers extend to the minimum and maximum values. The solid line represents the median estimate for each study, and individual data points denote any outliers. Similarly, in Figure A3, we present a forest plot of partial correlation coefficients across primary studies, revealing significant heterogeneity both between and within studies.

Figure 2: Variation of calculated PCCs across countries



Notes: The figure shows a box plot of PCCs reflecting the estimated relationship across countries.

In order to gain preliminary insights into the heterogeneity within our dataset, we examine mean values of partial correlation coefficients across different categories, as presented in Table 2. While we provide unweighted and weighted mean values, our ensuing discussion centres on the more dependable weighted mean values.

Studies that employ cross-sectional datasets for their analysis tend to yield higher mean positive effects than those using panel datasets. We also observe significant variations between estimates reported in published and unpublished studies, suggesting the potential presence of publication bias within the collected estimates. Notably, studies that overlook the issue of

endogeneity in their analysis of the impact of gender diversity on financial performance produce significantly lower weighted means compared to studies that effectively account for endogeneity. Regarding the mean effect across different estimation methods, we notice minimal differences among the techniques, except for the Random Effects approach with mean 0.0637 and Fixed Effects approach with mean 0.016.

Comparing the origins of the data by country, Table 2 reveals substantial variation. For example, studies conducted in South and Central America showed a negative mean of -0.0124, while European countries exhibit the highest mean effect of 0.0518 across all spatial specifications. Further, studies examining non-emerging economies or focusing on the financial industry yield more pronounced positive results.

When examining different specifications of leadership positions, we find noticeable differences in the mean effect estimates, ranging from 0.0212 to 0.0414. Further, studies using the gender diversity index or the number of female leaders report considerably lower positive effects than those using the proportion of female leaders or dummy specifications. Additionally, studies considering the critical mass of females in the upper echelon yield lower mean estimates than those that do not incorporate this factor. In our dataset, financial measures report substantial differences, with accounting means equal to 0.0269, market measures equal to 0.0409, and other measures equal to 0.0607.

In addition, we integrate a variety of distinct corporate and leadership attributes that have an impact on the relationship under scrutiny. Particularly, the duality of CEOs and the tenure of leaders exhibit notably elevated PCCs in comparison to other characteristics. Conversely, primary studies that incorporate the age of leaders tend to drive the mean towards negative estimates.

2.4 Variable Measurement

In the following subsections we focus on how are female leadership and financial performance measured in primary studies.

Table 2: Partial correlation coefficients for distinct subsets

		Weighted			Unweighted		
Variable	n	Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI
Data Type							
Cross-sectional data	1,131	0.0409	-0.0154	0.0279	0.0055	-0.0065	0.0174
Panel data	1,131	0.0311	0.0174	0.0359	0.0265	0.0212	0.0319
Publication characteristics							
Unpublished	1,131	-0.0185	-0.0300	0.0028	-0.0128	-0.0295	0.0039
Published	1,131	0.0357	0.0150	0.0333	0.0237	0.0186	0.0287
Endogeneity							
No endogeneity control	1,131	0.0288	-0.0033	0.0300	0.0124	0.0039	0.0210
Partial endogeneity control	1,131	0.0333	0.0142	0.0380	0.0252	0.0178	0.0326
Full endogeneity control	1,131	0.0410	0.0160	0.0459	0.0305	0.0211	0.0400
Estimation Method							
Elementary approach	1,131	0.0307	0.0023	0.0316	0.0155	0.0075	0.0236
Fixed effects	1,131	0.0160	0.0047	0.0234	0.0133	0.0069	0.0196
Random effects	1,131	0.0637	0.1160	0.0981	0.0506	0.0255	0.0756
Multi equation approach	1,131	0.0304	0.0207	0.0483	0.0346	0.0235	0.0457
Generalised methods of moments	1,131	0.0488	0.0111	0.0571	0.0340	0.0200	0.0481
Matching	1,131	0.0241	-0.0074	0.0800	0.0372	-0.0006	0.0749
Spatial Variation							
Cross country	1,131	0.0478	0.0278	0.0601	0.0429	0.0333	0.0525
USA	1,131	0.0061	-0.0026	0.0298	0.0139	0.0041	0.0237
South and Central America	1,131	-0.0124	-0.0132	0.0209	0.0033	-0.0075	0.0141
Europe	1,131	0.0518	0.0117	0.0488	0.0293	0.0171	0.0416
Asia	1,131	0.0202	-0.0047	0.0300	0.0126	0.0040	0.0211
Africa	1,131	0.0270	-0.0824	0.0424	-0.0226	-0.0691	0.0240
Analytical Design							
Emerging market	1,131	0.0177	-0.0033	0.0270	0.0112	0.0038	0.0185
Non-emerging market	1,131	0.0404	0.0199	0.0398	0.0295	0.0230	0.0360
Financial industry	1,131	0.0545	0.0197	0.0580	0.0376	0.0258	0.0495
Non-financial industry	1,131	0.0313	0.0109	0.0296	0.0205	0.0152	0.0258
Specifications of leadership positions							
Chief executive officer	1,131	0.0414	-0.0087	0.0410	0.0165	0.0039	0.0291
Top management team	1,131	0.0212	-0.0178	0.0184	-0.0004	-0.0097	0.0088
Board of directors	1,131	0.0349	0.0191	0.0393	0.0284	0.0223	0.0345
Specifications of gender diversity							
Proportion	1,131	0.0451	0.0123	0.0382	0.0239	0.0166	0.0313
Number	1,131	0.0140	0.0290	0.0987	0.0602	0.0368	0.0836
Dummy	1,131	0.0226	0.0039	0.0305	0.0178	0.0106	0.0250
Index	1,131	0.0095	-0.0247	0.0300	0.0035	-0.0189	0.0259
Criticalmass considered	1,131	0.0222	0.0076	0.0340	0.0208	0.0112	0.0304
Criticalmass not considered	1,131	0.0340	0.0133	0.0320	0.0225	0.0172	0.0279
Financial measures							
Accounting-based	1,131	0.0269	0.0017	0.0239	0.0129	0.0077	0.0182
Market-based	1,131	0.0409	0.0225	0.0557	0.0383	0.0276	0.0490
Other	1,131	0.0607	0.0162	0.0747	0.0444	0.0248	0.0639
Control Variables							
Firm size	1,131	0.0336	0.0098	0.0284	0.0181	0.0130	0.0231
Team size	1,131	0.0352	0.0089	0.0293	0.0192	0.0133	0.0251
Independent	1,131	0.0385	0.0126	0.0366	0.0246	0.0176	0.0315
Duality	1,131	0.0631	0.0202	0.0540	0.0357	0.0268	0.0447
Age	1,131	-0.0022	-0.0183	0.0155	-0.0008	-0.0129	0.0113
Tenure	1,131	0.0510	0.0024	0.0589	0.0267	0.0110	0.0425
Prior performance	1,131	0.0233	0.0072	0.0503	0.0280	0.0148	0.0412

Notes: The table presents the mean values of PCCs across various data subsets. We provide both the weighted mean, being calculated by assigning weights based on the inverse of the number of estimates per study, and unweighted mean.

2.4.1 Measurement of Financial Performance

A company's financial performance indicates its capacity to generate revenue, manage its assets and liabilities, and cater to the economic interests of its shareholders and stakeholders (Investopedia, 2023b). Academic studies often employ measures such as return on assets (ROA), return on equity (ROE), or market-based ratios like Tobin's Q to gauge financial performance. However, the choice of these metrics can vary based on the research design and data availability, leading to the use of other ratios like return on investment (ROI) (Miller and Triana, 2009; Shrader et al., 1997); return on sales (ROS) (Isidro and Sobral, 2015); market-to-book (MB) (Bonn et al., 2004); and Sharpe ratio (Robb and Watson, 2012).

Approximately 65% of the estimates in the primary studies use ROA as the financial performance proxy. ROA assesses the efficiency of a company in generating revenues in excess of actual expenses from a given portfolio of assets. It is computed as Net Income/Total assets. ROE is employed as a dependent variable in around 27% of primary studies. It is determined as Net Income/Shareholder's Equity. This ratio evaluates the returns generated for the company's shareholders (Investopedia, 2023d). ROA and ROE are accounting-based measures that look backwards in time. They are based on the financial performance that a company has reported in the recent past (Haslam et al., 2010). Robb and Watson (2012) argue that ROA is a more reliable measure than ROE as ROE can be further decomposed into the product of ROA and leverage. When a company increases its leverage and debt, the ROE will be higher in comparison to the ROA. They claim that the level of debt depends solely on the specific company, and hence the ROA is a more objective measure.

Tobin's Q is a financial performance measure that reflects a firm's ability to create value by dividing its total market value by its total asset value (Investopedia, 2023c). This measure is used in 52% of the primary research. It is considered more reliable than accounting-based measures as it reflects the future potential of a firm's performance (Haslam et al., 2010). Furthermore, it is a standardised measure with intuitive interpretation, with a ratio greater than one indicating a high competitive advantage for the firm. Unlike ROE and ROA, Tobin's Q is based on objective data, not self-reported data (Lindenberg and Ross, 1981), thereby eliminating accounting convention bias.

2.4.2 Measurement of Female Leadership

Several proxies are utilised in the literature to express gender diversity in upper-level management, including the proportion of females in top positions, dummy variables, or indices.

The proportion of females in leadership roles is a widely accepted and well-established measure of gender diversity (Robb and Watson, 2012; Strøm et al., 2014). It offers a clear and quantifiable representation of female involvement and can be displayed as a fraction, ratio, or percentage. This metric allows for easy comparison across companies and industries and tracking changes over time.

Research examining the presence of women in leadership roles often employs dummy variable analysis. This method assigns a value of 1 to indicate the presence of at least one woman and 0 otherwise (Trinh et al., 2018). However, more than mere presence is required to understand the relationship between female leadership and financial performance. If the data permits, researchers apply the critical mass theory and set a higher threshold for female representation, i.e., the dummy equals 1 if there are at least three women in the group and 0 otherwise (Erkut et al., 2008).

The Blau's Index (1997) is a widely adopted metric for quantifying diversity within categorical variables and has been recommended as a suitable method by experts in the field of diversity research (Harrison and Klein, 2007). Blau's index serves as an impartial metric for measuring diversity among gender. In the context of gender diversity, Blau's Index varies from 0 (when a single gender is represented) to 0.50 (when there is a parity of representation between genders).

The Shannon Index (2001) is a well-established diversity index commonly cited in the literature. It shares similarities with Blau's Index but is always higher and more sensitive to small differences in gender composition (Campbell and Minguez-Vera, 2008). The index ranges from 0 (when members exclusively belong to male or female groups) to 0.6931 (when both genders are equally represented).

3 Publication Bias

Previous sections have explored theoretical frameworks that delineate the relationship between female leadership and financial performance. Further, we summarised collected estimates to identify the average effect and prominent empirical trends. Nevertheless, Ioannidis et al. (2017) report that even though simple weighted or unweighted averages of all reported estimates can aid in eliminating sampling errors and random misspecification bias when sufficient estimates are available, they fail to shed light on the credibility of the results, thereby potentially leading to biased summary effects.

Historically, the trustworthiness of specific findings has been assessed by journal editors and peer reviewers, who hold the ultimate authority on which outcomes are published. Even with attempts to standardise this decision-making procedure, journal editors often predispose towards results with particular characteristics. Publication bias, also known as the "file drawer problem" (Rosenthal, 1979), has been a perennial concern for meta-analysts due to its propensity to skew results and influence the perceived scale of empirical effects. To counter this issue, researchers have incorporated working papers and other unpublished materials in their studies, regardless of their significance (Sterling, 1959). Card and Krueger (1995) identify three primary sources of publication bias in economics: an inclination amongst editors and reviewers towards papers that resonate with conventional viewpoints, the selection of models by researchers based on anticipated results, and a common propensity among both reviewers and researchers to favour statistically significant findings.

Publication bias does not necessarily stem from deliberate actions. Authors may prefer to submit significant findings based on the logical assumption that such studies are more likely to be accepted. In contrast, referees and editors might favour significant results, believing they convey more valuable information. Irrespective of the intention, this trend contributes to the underrepresentation of insignificant findings and skewed evaluations in the scholarly literature (Stanley, 2005).

To further our understanding of systematic misreporting, we investigate the possibility that the existing empirical evidence on gender diversity in corporate settings is influenced by publication selection bias. An individual with a surface-level grasp of the subject might dismiss the likelihood of such bias for two reasons. Firstly, the coexistence of multiple theories which

compellingly advocate for both positive and negative outcomes could reduce the possibility of research findings being disproportionately influenced by a widely accepted theory. Secondly, the apparent lack of consensus among scholars regarding the financial implications of gender diversity within corporations suggests a reduced bias towards specific findings.

Prior meta-analyses have acknowledged the potential presence of selection bias, but their treatment of the issue has been relatively limited. Only Pletzer et al. (2015) directly confront the problem by employing a funnel plot and fail-safe N test, ultimately finding no substantial evidence of publication bias.

If publication bias impacts compiled data, the descriptive statistics presented in Section 2 (Table 2) could be skewed, failing to represent the true mean effect accurately. Therefore, in this section, we assess the trustworthiness of empirical evidence regarding the impact of female leaders on financial outcomes by investigating the influence of selection bias on the collected estimates through the application of funnel plot, as well as linear and non-linear tests.

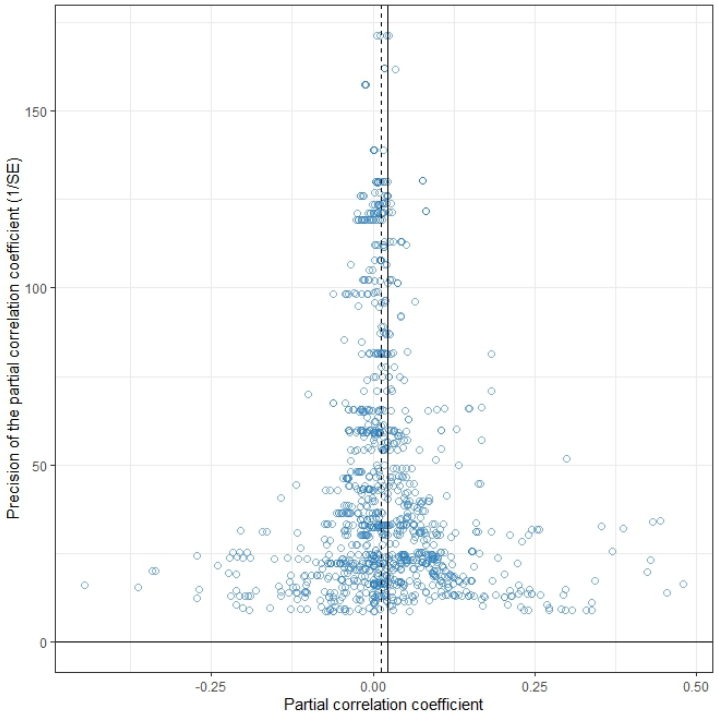
3.1 Funnel Plot

We initiate our examination of potential publication bias by adopting a visual evaluation instrument known as a Funnel Plot, as suggested by Egger et al. (1997). This straightforward technique generates a scatter plot, which displays the magnitude of estimated effects (in our case, partial correlation coefficients) on the horizontal axis versus their precision, defined as the inverse of the estimated standard errors, on the vertical axis. In the event of no publication bias, the scatter points should be symmetrically arranged around the average value line, forming an inverted funnel shape (Sterne et al., 2005). This symmetry indicates that the average value signifies the true effect, with the most accurate estimates situated near the average at the top of the graph and the least precise estimates widening the funnel at the base (Sterne and Harbord, 2004).

In contrast, an asymmetric funnel plot hints at the existence of publication bias, such as a preference for positive or negative estimates or a higher occurrence of statistically significant results. A hollow funnel shape signifies an underrepresentation of statistically insignificant values, regardless of the direction of effect (Stanley, 2005). However, it is crucial to acknowledge that asymmetry in the funnel plot could result from factors other than publication bias, such as

methodological disparities, data irregularities, or genuine heterogeneity among studies (Egger et al., 1997). To address these concerns, we conduct statistical tests for funnel asymmetry, as discussed in the subsequent sections.

Figure 3: The funnel plot



Notes: The diagram illustrates the funnel plot of partial correlation coefficients. Although we employ unwinsorised data to create this chart, we utilise winsorised data for quantitative analyses.

The diagram in Figure 3 hints at the absence of publication bias, as it mirrors the expected inverted funnel shape with a reasonable degree of symmetry. We even observe less precise estimates clustered at the bottom of the funnel. Importantly, across varying levels of precision, the diagram is not hollow. Upon closer examination, it becomes evident that our estimates, though seemingly symmetrically positioned around a value close to zero, actually exhibit a slight right skew, underscoring an underrepresentation of negative estimates. Our premature visual observation aligns with the result obtained by Pletzer et al. (2015), suggesting little evidence of publication bias as their funnel plot is symmetrical around the central vertical line.

While our visual assessments provide some insight, the subjective nature of such analysis calls for more stringent methods to ascertain the presence and magnitude of publication bias in the existing research. Recognising that a mere funnel plot cannot conclusively identify

publication bias, we conduct additional analyses in the following subsections.

3.2 Linear and Non-linear Tests

We proceed with our investigation by using a linear approach towards publication bias, the funnel asymmetry test (FAT), as proposed by Card and Krueger (1995). This test involves estimating the subsequent regression:

$$PCC_{is} = \beta_0 + \beta_1 SE(PCC_{is}) + \epsilon_{is}. \tag{7}$$

In this context, PCC_{is} and $SE(PCC_{is})$ denote the computed partial correlation coefficients and their corresponding standard errors, while ϵ_{is} signifies the error term, β_0 represents effect beyond bias, and the coefficient β_1 reflects the degree of publication bias, providing insights into its presence, direction, and magnitude.

Equation 7 measures the symmetry of the funnel plot (Egger et al., 1997). Essentially, FAT examines the correlation between reported estimates and their standard errors. In the case of publication bias, this correlation should be significant. Conversely, in the absence of publication bias, the test identifies a zero correlation (Stanley, 2005). From the perspective of hypothesis testing, the null hypothesis represents no publication bias, i.e., H_0: $\beta_1 = 0$, against the alternative of publication bias being present, H_1: $\beta_1 \neq 0$.

Doucouliagos and Stanley (2013) classify the degree of publication selectivity as *little to modest* when the FAT test is not statistically significant or if the absolute value of β_1 is less than 1. They consider it *substantial* if the FAT test is statistically significant and the absolute value of β_1 lies between 1 and 2, and they describe it as *severe* when the FAT test is statistically significant and the absolute value of β_1 exceeds 2.

We estimate Equation 7 using the following model specifications. Initially, we use ordinary least squares. Then, we examine publication bias by estimating the model with between-study variance. We do not utilise within-study variance in our meta-analysis due to dataset imbalance and some primary studies providing only one effect estimate. Subsequently, we weight the estimates by the inverse of the number of observations per study - a method commonly applied in meta-analyses to ensure that studies with varying reported observations are given equal weight (Gechert et al., 2022). Our final model specification considers the potential is-

sue of heteroskedasticity, as the variance of the outcome variable is captured directly by the explanatory variable (Stanley, 2005). We address this concern and enhance efficiency by adopting a standard practice in meta-analysis - weighting the equation by precision (Stanley and Doucouliagos, 2017).

Lastly, we tackle heteroskedasticity in FAT-PAT by clustering standard errors at the study level. As previously mentioned, the assumption of independently and identically distributed error terms ($\epsilon_{is} \sim$ iid) might not be valid. To avoid erroneous results and inferences, we cluster standard errors at the study level, accounting for possible correlations within the same studies and independence between different studies. While our sample fulfils the minimum criteria for reliable inference, the uneven cluster sizes may still introduce bias (MacKinnon and Webb, 2017). To remedy cluster imbalance, we apply the wild bootstrap clustering method as advised by Gechert et al. (2019) and provide the 95% confidence interval for all specifications, except for the between-effect estimation at the study level.

The results from the various model specifications are summarised in Table 3. Three out of four estimations show positive and significant publication bias at the 5% level. According to the classification of Doucouliagos and Stanley (2013), this bias is categorised as little to mild. When we adjust for the uneven distribution of estimates across studies, the detected publication bias loses statistical significance - suggesting that some studies might be driving the publication bias.

When comparing values of the estimated mean beyond bias with the weighted and unweighted means from the primary studies, which are 0.033 and 0.022, respectively (as detailed in Table 1), it is evident that the bias-corrected mean is substantially lower, ranging from 0.01 to 0.0229. This observation implies an influence of publication bias on the PCCs, suggesting that the ultimate impact of female leadership on financial outcomes remains vague.

While funnel and precision asymmetry tests offer valid methodologies for identifying publication bias and estimating effect size after bias adjustment (Stanley, 2008), they rely on a presumed linear association between partial correlation coefficients and standard errors. However, this presumption may not hold, resulting in potentially misleading results. Stanley et al. (2010) underscore that this assumption of linearity may not always hold true, especially for highly accurate estimates located at the peak of the funnel plot, which are less susceptible to

Table 3: Linear and non-linear techniques to detect publication bias

Panel A: linear tests	OLS	Between	Weighted by study	Weighted by precision
SE	0.2671*	0.2965***	0.2604	0.3286***
(Publication bias)	(0.1345)	(0.0998)	(0.3513)	(0.0998)
	[-0.527;1.131]		[-0.662;1.191]	[-0.253;0.894]
Constant	0.0123**	0.0114***	0.0229*	0.01***
(Mean Beyond Bias)	(0.0041)	(0.0036)	(0.0117)	(0.0024)
	[-0.015;0.036]		[-0.007;0.052]	[-0.006;0.023]
Observations	1,131	1,131	1,131	1,131
Studies	96	96	96	96
Panel B: non-linear tests	STEM method	WAAP	Selection Model	Endogenous Kink
Effect Beyond Bias	0.0187**	0.0121***	0.021**	0.009***
	(0.0081)	(0.0019)	(0.004)	(0.002)
Observations	1,131	1,131	1,131	1,131
Studies	96	96	96	96
Panel C: relaxing exogeneity assumption	IV	IV)		
	1/sqrt(sample size)	log(sample size)		
SE	0.2938**	0.3233**		
(Publication bias)	(0.1378)	(0.1135)		
	[0.0298;0.5623]	[0.1049;0.5438]		
Constant	0.0113***	0.0102***		
(Mean Beyond Bias)	(0.0041)	(0.0029)		
	[0.0033;0.0192]	[0.0044;0.016]		
F-statistics	241,929.79	6,992.186		
Observations	1,131	1,131		
Studies	96	96		

Notes: Panel A presents linear approaches for addressing publication bias. Panel B presents the outcomes of non-linear analysis, demonstrating the size and significance of the impact after adjusting for publication bias. Panel C showcases approaches for addressing publication bias while relaxing exogeneity assumption. IV refers to regressions that use the inverse of the square root of the number of observations and the logarithm of sample size as an instrument. WAAP stands for Weighted Average of the Adequately Powered, standard errors are in the parentheses and 95% confidence intervals from wild bootstrap in square brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

publication bias due to their small standard errors. Consequently, the reliance on linear approximation may overstate the presence of publication bias, thereby underestimating the actual underlying effect. As shown in recent meta-analyses, we use these alternative non-linear tests to establish the empirical effect that is independent of publication bias (e.g., cazachevici et al., 2020; Matousek et al., 2019). Namely Weighted Average of Adequately Powered (WAAP) strategy, formulated by Ioannidis et al. (2017), Selection Model proposed by Andrews and Kasy (2019), we also employ the STEM-based method and the Endogenous Kink method introduced by Furukawa (2019) and Bom and Rachinger (2019), respectively.

To compute the WAAP and STEM-based estimates, we utilise R Studio. Meanwhile, for the Selection Model, we employ their online application, specifying a commonly accepted t-statistic threshold of 1.96 and predicting the effect distribution. Finally, the Endogenous Kink estimate is derived via Stata.

The results from non-linear tests presented in Table 3 align with findings from the FAT-PET

tests. Bias-corrected effects, range from 0.009 to 0.021. Although these effects are statistically significant, they remain relatively low and fall below the means of both weighted and unweighted partial correlation coefficients, thereby affirming our previously observed patterns.

Previous tests confirm positive publication selection in studies. Nevertheless, the results might be biased due to endogeneity. Stanley (2005) asserts that bias in estimating Equation 7 could arise from random sampling errors and the simultaneous computation of effect estimates and standard errors, both of which could be influenced by the estimation method of the primary study. He recommends using the sample size as an instrumental variable to address this issue. The sample size is expected to meet the necessary conditions because it correlates with standard errors as larger-sample studies generally produce smaller standard errors, and it is unlikely, though not entirely improbable, that it is correlated with the selected estimation method.

Accordingly, we implement two instruments: the inverse of the square root of the sample size of primary studies (Gechert et al., 2022) and the logarithm of the sample size. The robustness of the instrumental variables is evaluated using the weak instruments test, which produces an F-statistic significantly surpassing the conventional benchmark value of 10, thereby affirming the strength of the instruments.

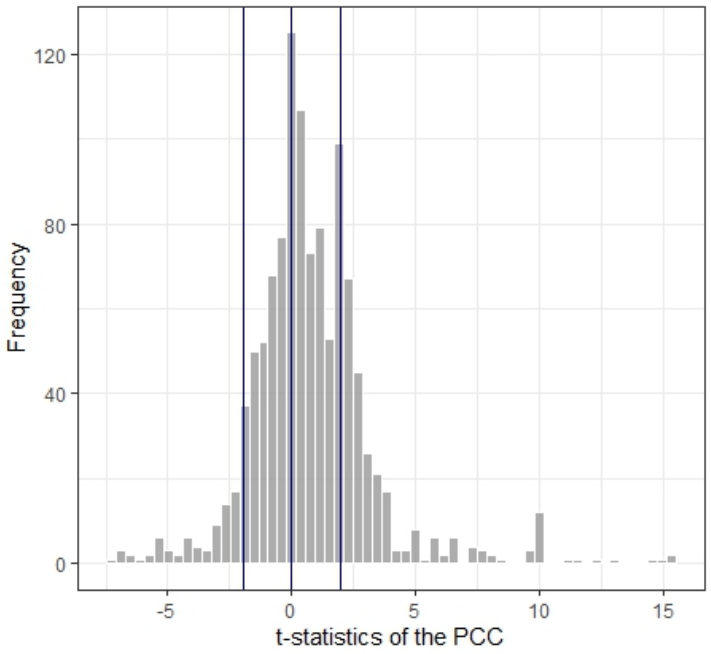
The findings, presented in Panel C in Table 3, align with conclusions drawn from preceding sections, revealing a negligible effect of female leadership on financial performance and a modest degree of publication bias. The publication bias is significant at the 5% level. The values of mean beyond bias are significant at 5% and 10% level, however, compared to the weighted and unweighted mean of PCCs, they are substantially lower.

3.3 Caliper Test

In previous sections, we delved into publication bias by examining the relationship between partial correlation coefficients and their corresponding standard errors. To wrap up our bias investigation, we employ the caliper test, a method introduced by Gerber and Malhotra (2008). This test is a prevalent tool in modern meta-analyses, as demonstrated in the studies of Havranek et al. (2020) or Matousek et al. (2019). This test compares the number of estimates located within equal-sized ranges above and below a specific t-statistic threshold, known as a caliper. This approach offers two key advantages: it enables direct comparison of the collected or cal-

culated t-statistics derived from standard errors and addresses potential endogeneity, similar to the instrumental variable estimation discussed in Section 3.

Figure 4: The distribution of t-statistics



Notes: The figure represents the distribution of t-statistics of the reported estimates of the elasticity overlaid on a corresponding normal distribution. Red lines represents critical value of 1.96 associated with significance at the 5% level and the value of 0 associated with changing the sign of the estimate. We exclude estimates with large t-statistics from the figure for ease of exposition but include them in statistical tests.

Figure 4 illustrates the distribution of the t-statistic of the collected estimates. For our analysis, we adopt the t-statistic of ± 1.96, which is commonly used. In the absence of publication bias, the likelihood of observing a result just above or below the thresholds is expected to be equal, assuming that the distribution generating a coefficient estimate is a continuous probability distribution. If the contrary is true, these critical values significantly influence which findings are published, indicating the existence of publication selection bias.

The results of the caliper test, as outlined in Table 4, show no clear publication selection bias between significant and non-significant negative estimates. For the t-statistic equal to 1.96, positive significant estimates surpass positive non-significant estimates yielding a proportion of 68% and 32% within the narrowest 5% band. This insight enhances our analysis, suggesting that researchers tend to prefer positive estimates to negative ones, as corroborated by linear tests,

and they also lean towards significant positive estimates rather than non-significant positive findings.

Table 4: Caliper Test

	Treshold = -1.96	Treshold = 1.96
Caliper width 0.05	0.5625***	0.6786***
	(0.1281)	(0.0898)
Caliper width 0.1	0.5***	0.5556***
	(0.109)	(0.0683)
Observations	1,131	1,131
Studies	96	96

Notes: The table presents the outcomes of the caliper test which is conducted for critical values -1.96 and 1.96 and the results are displayed in the separate columns. Standard errors are provided in parentheses. Statistical significance is reported as follows * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In summary, the majority of tests conducted in this chapter convey a consistent narrative. These tests reveal the presence of positive publication bias in the literature, with bias estimates ranging from 0.2604 to 0.3286 and minimal effect of female leaders on the financial outcomes of companies. While not all of these results reach statistical significance, it is essential to note that no statistically significant results suggest a contrary finding, such as a negative publication bias or mean beyond bias. However, these findings warrant further examination, as they may be covertly correlated with unidentified factors.

4 Heterogeneity

Investigations into the impact of female leadership on financial performance reveal significant variation within the literature, even when partial correlation coefficients serve as the basis of comparison (as shown in Figure A3). This section explores additional sources of heterogeneity, accounting for context-specific aspects of each study. Beyond publication bias, the variation could stem from methodological differences and data used (Adams, 2016). We aim to pinpoint those attributes which most significantly influence the relationship between gender diversity and company financial performance. Furthermore, we seek to discern whether the relationship between partial correlation coefficients and their standard errors remains solid after adjusting for other variables. If this relationship endures, it can act as a robustness test for publication bias (Havranek et al., 2021).

Previous research underlines the significance of addressing heterogeneity. Reddy and Jad-

hav (2019) emphasise external factors affecting board gender diversity, such as firm size, board size, industry, ownership type, customer base, and sociocultural factors. They also note limited evidence connecting directors' personal traits to gender diversity in top management. Consequently, studies examining the impact of board gender diversity on firm performance yield inconclusive results.

Meta-analyses written on the topic provide the following insights into data variability. Eagly and Carli (2003) address the variability of estimates by examining unweighted and weighted mean values for study-level effect sizes. They focus on transformational, transactional, and laissez-faire leadership styles, providing a nuanced understanding of heterogeneity in the field.

Post and Byron (2015) adopt a multifaceted approach in their meta-analysis. They commence with random effects analyses, leveraging Wilson's meta-analysis macros for SPSS to accommodate the diversity of the studies. Next, they calculate the mean effect size across all studies and establish 95% confidence intervals, providing an overall measure of the effect's strength and direction and the range within which the true effect size likely falls. They also use the Q Statistic to assess the variability in effect sizes across studies, which can suggest the presence of moderating variables. Lastly, they undertake a meta-regression to ascertain if any continuous moderator variables (i.e. shareholder protection strength and gender parity) account for the variability in effect sizes.

Pletzer et al. (2015) implement sensitivity and moderator analyses in their study. They start with a cumulative analysis, sequentially integrating studies to evaluate the stability of the overall mean effect size, thereby providing insights into the robustness of their findings. They then apply a one-study removed analysis to gauge the influence of individual studies on the overall mean effect size, helping to identify studies with unusually significant or insignificant impacts on the results. Finally, they perform moderator analyses, including subgroup analyses and univariate meta-regressions, to explore the effect of systematic differences between studies with moderators for mean board size, country development and income.

In their 2018 study, Hoobler et al. utilise a trio of methodological strategies to manage result variability. The first strategy, correction of study artefacts, involves assigning a reliability coefficient of 0.8 to both dependent and independent variables when not provided by the original studies. They also adjust for sample size, giving more significance to more extensive

studies. The second strategy, calculation of effect sizes, involves averaging correlations across multiple performance measures reported in the studies. Further, they compute 95% confidence intervals to ascertain if the mean effect size significantly differs from zero. The final strategy, moderator analysis, pinpoints factors potentially influencing the relationship between the dependent and independent variables. They employ an 80% credibility interval, compute the Q Statistic, and test potential moderators (i.e. country-level gender egalitarianism, board meeting frequency/activity, and board size) using weighted least squares regression. Each of the referenced meta-analyses reveals a positive, albeit minimal, impact on the relationship under scrutiny.

In this chapter, we embark on a more profound exploration of the sources of variation - our attention pivots towards the elements of study design that contribute to the observed heterogeneity in PCCs. We provide a comprehensive description of the variables collected to encapsulate systematic differences. Subsequently, we clarify the employment of Bayesian Model Averaging, Frequentist Model Averaging, and frequentist check techniques as instruments for investigating heterogeneity. The ensuing discussion is dedicated to the presentation of our findings and providing a thorough understanding of variation sources.

4.1 Explanatory Variables

Data Characteristics Beyond the conventional effect estimate and its corresponding standard error, we include four additional variables depicting unique features of datasets used in primary studies. The employed sample sizes exhibit notable variation. For instance, Dezsö and Ross's (2012) nationwide dataset, with its 21,790 data points, greatly contrasts with Vu's et al. (2019) study, which involved only 84 observations from Vietnam. Such restricted datasets could obscure or understate the influence of women's leadership on financial performance. This underscores the need for large-scale, representative datasets to encapsulate the relationship being examined precisely. To account for sample size disparities, we introduce a variable, labelled *Sample size*, that essentially denotes the logarithm of the observations in use.

While some scholars depend purely on cross-sectional data, others have opted for panel datasets. Panel data can offer insights into causality and control unobserved individual heterogeneity that remains invariant over time, thus tackling potential endogeneity issues Naseem et

al., 2019. To reflect this methodological choice, we integrate a binary variable, *Panel*, into our examination.

We also include variable *Avg. data year* designed to standardise the year when the data was utilised. It is based on the presumption that the estimated influence of women's leadership on financial performance might vary across different years due to changing societal norms and policies.

Finally, within this assortment of variables, we incorporate a variable to capture the number of variables included in the regression model - the origin of the estimated effect of gender. This attribute may affect the heterogeneity observed in the study outcomes. The number of variables can significantly vary both within and across different studies. It extends from a bare minimum of one (as seen in Adams, 2016 or Rose, 2007) to an extraordinary maximum of over 2,500, including dummies of all firms and years into the model (as in Dezsö and Ross, 2012). This extensive variance underscores the potential influence the variable *No of variables* might have on the heterogeneity of results.

Publication Characteristics The initial variable in this set pertains to the *Publication year*, computed by taking the logarithm of the year of the primary study's publication, adjusted by subtracting 1997 - the year of publication of the earliest study in our sample. As the oldest study in our sample was published in 1997 (Shrader et al.) and the most recent in 2023 (Gull et al.), we anticipate this temporal variance could introduce heterogeneity in primary study outcomes.

The following characteristic relates to *Citations*. This parameter corresponds to the logarithm of the total citations a primary study received, normalised by the number of years since its publication. This adjustment prevents older studies from unjustifiable weighting due to high citation counts arising from longevity rather than merit.

As our meta-analysis focuses on gender diversity, we are interested in whether the author's gender influences the outcomes. Thus, we include *Female ratio*, a variable expressing the proportion of female authors to all study authors.

We further implement variable *Published* for studies published in the peer-reviewed journal. The last variable within this category is the *Impact factor*, a proxy for the quality of primary studies. The impact factor is based on the Recursive Impact Factor of IDEAS/RePEc (Re-

search Papers in Economics), initially introduced by Zimmermann (2013). Studies not listed or unpublished studies are assigned an impact factor of zero.

Estimation Methods The estimation techniques applied to investigate the subject are classified into five distinct variables: *Elementary approach* incorporating ordinary least squares, generalised least squares or generalised linear model estimator, *FE* indicating fixed effects estimation, *RE* symbolising random effects estimation, *Multi equation* for 2SLS and 3SLS estimators, and *GMM* for any GMM extensions employed.

A binary variable, *Matching*, is introduced to denote whether authors incorporate any form of matching, such as Propensity Score Matching (PSM) or Coarsened Exact Matching (CEM), in their analysis.

We anticipate that the selected estimation method, each with its distinct assumptions and capabilities to account for unobserved differences, may influence the reported estimates. This hypothesis corresponds with the findings of Bennouri et al. (2018), suggesting that different estimation methods, like OLS, FE, and IV, can yield diverse estimates even when applied to the same datasets.

Analytical Design We acknowledge the susceptibility of the scrutinised relationship to a range of endogeneity biases. To address these, we classify our studies into three categories using binary variables: *Endogeneity* denotes a study overlooking all biases; *Partial endogeneity* signifies authors acknowledging and rectifying some biases; *Full endogeneity* corresponds to a study that successfully tackles all biases. The underlying rationale for this separation lies in the inclination of numerous studies to adopt an intermediary approach. They aim to demonstrate the evolution of results in a hierarchical manner while effectively mitigating endogeneity concerns within their datasets (Adams, 2016). This sequential approach allows us to perform a robustness check further, exclusively focusing on studies that exhaustively employ all available methods to establish closer proximity to causal relationships.

We also examine the geographical scope of the primary studies. While a significant percentage of studies, 21.9%, are derived from European datasets, and 16.7% pertain to observations from the U.S. Our dataset represents other geographical regions as well, i.e. studies conducted in Asia (25%), Africa (4.2%), Australia (4.2%), South and Central America (4.2%), and the

rest employ global datasets. To account for variations in corporate governance and workforce attributes across nations, we classify studies based on the geographical origin of their data and introduce corresponding binary variables such as *Europe*, *USA*, *Asia*, *Africa*, *Cross country* for global datasets and variable *Other* incorporates the rest of the regions.

The distinction between financial and non-financial sectors is crucial due to their disparate natures (Horváth et al., 2012; Kweh et al., 2019). Financial institutions are subject to strict regulations and oversight, thus cultivating a more conservative, risk-averse environment that could impact women's roles in leadership (Julizaerma and Sori, 2012). Hence, we establish a variable, *Financial*, to highlight studies focusing solely on the financial industry (Lafuente and Vailland, 2019) or including financial institution sub-samples (Vu et al., 2019).

The business landscape, societal norms, and regulations vary considerably between developed and emerging markets (Terjesen and Singh, 2008). While marked by rapid growth and promising opportunities, emerging markets often have less developed institutional infrastructures, weaker legal systems, and less stable economic-political environments (Investopedia, 2023a). These elements can affect business operations and women's leadership roles. We introduce the variable *Emerging* to address possible heterogeneity, signifying studies conducted in an emerging market context.

Variable Specifications Researchers adopt diverse methods to quantify the financial outcomes of the company and the leadership status of women in top positions. Some studies employ the numerical count or proportion of female leaders, while others favour diversity indices or dummy variables. To accommodate the potential variation arising from these diverse methods of characterising gender diversity, we formulate binary variables: *Number*, *Proportion*, *Index* and *Dummy*.

Some of the authors consider critical mass theory in their analysis. As per Joecks et al. (2013), it is essential to distinguish between females who serve merely as token representatives and those who contribute to creating a critical mass within the leadership team. We construct a binary variable, *Criticalmass*, for models incorporating this perspective.

In order to capture potential variances in focus on different roles within corporations, we include three variables reflecting female leaders as part of top management teams, boards of directors, or serve as CEOs. Although each of these roles implies a high-ranking status, the

relative focus on them across different studies could generate heterogeneous results. Hence, we formulate distinct variables: *TMT*, *Board*, and *CEO*.

The financial outcomes of a company are represented by three variables: *Accounting*, which corresponds to accounting measures, and *Market*, which pertains to market-based measures. However, eight primary studies employ unique dependent variables that do not resonate with either of these measures. In such instances, we construct a dummy variable, *Other*. For example, Nadeem et al. (2020) utilise the variable ECONO, which signifies the total scores on a firm's financial performance. Similarly, Strøm et al. (2014) use financial self-sufficiency as a proxy for the company's financial performance.

Control Variables In our analysis, we also integrate several key characteristics pertaining to female leaders, leadership teams, and firms. Among a broad array of potential factors, we have chosen to concentrate on the size of the firm, the size of the leadership team, the independence of the leaders, the dual role of CEOs, and the age and tenure of the leaders, in addition to the prior performance of the company. These factors were identified as the most significant in our primary research so we include the dummy variables *Firm size*, *Team size*, *Independent*, *Duality*, *Age*, *Tenure*, and *Prior performance* in cases where these specifications are referenced in the primary analysis.

4.2 Estimation

Introducing BMA We have gathered 45 distinctive study features, that are outlined in Table 5, covering various aspects of the study design. Our aim is to identify the features that consistently impact the relationship coefficients between female leadership and firm financial performance. Adopting a direct approach involves regressing the calculated partial correlation coefficients on the entire range of supplementary explanatory variables. However, this may lead to imprecise results due to the potential inflation of standard errors caused by the extensive number of predictors. To address this, many empirical studies follow a sequential elimination process to build an optimal model free from irrelevant predictors. Yet, this method carries the risk of retaining an insignificant variable or discarding a crucial one. Koop (2003) highlights that the likelihood of such errors increases significantly with the growing number of models considered. Given our study, a selection procedure from 2^{45} models presents a daunting task.

Table 5: Description of variables

Variable	Description	Mean	SD
PCC	partial correlation coefficient	0.0224	0.0843
Standard error	standard error of PCC	0.0376	0.0251
Data Characteristics			
Panel	=1 if a study implement panel dataset	0.8019	0.3987
Avg year	= logarithm of average year of study	2.8594	0.3856
No of variables	= logarithm of explanatory variables included in a study	3.058	1.5733
Sample size	= logarithm of sample size	7.101	1.4740
Publication characteristics			
Publication year	= logarithm of the year a study was published	2.896	0.492
Citations	= logarithm of total citations of a study	3.106	1.2113
Female ratio	= proportion of the female authors of a study	0.4066	0.3289
Impact factor	= proxy for the quality of primary studies	0.1554	0.340
Published	=1 study was published in the peer review journal	0.9637	0.1870
Estimation methods			
Emelentary approach *(ref. group)*	=1 if estimated by OLS, GLM or GLS	0.3873	0.4873
FE	=1 if estimated by fixed effects model	0.2608	0.4393
RE	=1 if estimated by random effects model	0.0504	0.2189
Multi equation	=1 if estimated by 2SLS or 3SLS	0.0867	0.2814
GMM	=1 if estimated by any of GMM extensions	0.2149	0.4109
Matching	=1 if the analysis is done on matched sample	0.0610	0.2395
Analytical design			
Endogeneity *(ref. group)*	=1 if study does not control for endogeneity	0.3431	0.4749
Partial endogeneity	=1 if study partially controls for endogeneity	0.313	0.4639
Full endogeneity	=1 if study fully controls for endogeneity	0.3351	0.4722
Europe	=1 if a study used data from European country	0.2317	0.4221
USA	=1 if a study used data from the US	0.641	0.3305
Asia	=1 if a study used data from Asian country	0.2458	0.4308
Africa	=1 if a study used data from African country	0.0327	2.060
Cross country	=1 if a study used global dataset	0.2529	0.4349
Other *(ref. group)*	=1 if a study used data from other countries	0.1123	0.3081
Financial	=1 if study analyses companies within financial industry	0.1096	0.3126
Emerging	=1 if study analyses emerging market	0.3908	0.4881
Variable specifications			
Number *(ref. group)*	=1 if gender diversity is expressed by number of females	0.0592	0.088
Proportion	=1 if gender diversity is expressed by proportion of females	0.4359	1.588
Index	=1 if gender diversity is expressed by diversity index	0.0442	0.619
Dummy	=1 if gender diversity is expressed by dummy variable	0.4607	1.395
Criticalmass	=1 if critical mass is considered in the model	0.099	0.619
TMT *(ref. group)*	=1 if study analyses top management teams	0.1388	0.3459
CEO	=1 if study analyses chief executive officers	0.1724	0.3779
Board	=1 if study analyses board of directors	0.6879	0.4636
Accounting	=1 if accounting-based financial measures are considered	0.6384	0.4807
Market	=1 if market-based financial measures are considered	0.3218	0.4674
Other *(ref. group)*	=1 if other financial measures are considered	0.0398	0.1955
Control variables			
Firm size	=1 if size of the firm is included in analysis	0.8576	0.3496
Team size	=1 if size of team is included in analysis	0.7445	0.4363
Independet	=1 if independence of leaders is included in analysis	0.5455	0.4981
Duality	=1 if duality of CEOs is included in analysis	0.3324	0.4713
Age	=1 if age of leaders is included in anaylsis	0.1194	0.3244
Tenure	=1 if tenure of leaders is included in a study	0.1141	0.3180
Prior performance	=1 if past performance is included in a study	0.2308	0.4215

Note: The table outlines explanatory variables collected to analyse studied relationship. It further displays the mean and standard deviation of variables eligible for use in meta-regressions.

We utilise the innovative Bayesian Model Averaging (BMA) estimation approach to explore heterogeneity and comprehensively address the challenges of multiple potential models (Moral-Benito, 2012). Unlike conventional approaches that rely on a single model or consider all possible models, BMA thoughtfully incorporates diverse covariate subsamples (Kass and Raftery, 1995). By conducting multiple regressions using varying subsets of explanatory variables and skillfully combining them through a weighted average (Zeugner and Feldkircher, 2015), BMA exhibits superior predictive capabilities compared to traditional methods like sequential regression (Kass and Raftery, 1995). The relationship under scrutiny can be formulated as follows:

$$PCC_{is} = \beta_0 + \beta_1(X_{is}) + \beta_2 SE(PCC_{is}) + \epsilon_{is}. \tag{8}$$

Where PCC_{is} corresponds to partial correlation coefficients, β_0 is a constant term; X_{is} stands for additional independent variables, β_2 measures the intensity of publication bias, $SE(PCC_{is})$ is a standard error and ϵ_{is} refers to an error term.

Regarding the explanation of BMA and the ensuing analysis, we adopt the methodology proposed by Zeugner and Feldkircher (2015) and Hasan et al. (2018). Considering K variables, there exist 2^K potential combinations, resulting 2^K models M. The estimation technique encompasses the following components: Posterior Model Probability, Posterior Inclusion Probability, Posterior Mean, Posterior Variance

4.2.1 Implementing BMA

Moving forward, we apply the BMA method to estimate Equation 8, making use of the explanatory variables presented in subsection 4.1. However, several factors limit the integration of all gathered variables in the analysis. First, incorporating every binary variable from a given set risks generating a dummy variable trap. To avert this, we eliminate one variable in each category prone to perfect multicollinearity, denoted as the "reference category" in Table 5. Second, we assess the correlation matrix of the remaining variables to avoid including highly correlated variables. We observe correlation of -0.93 between variables *Standard error* and *Sample size*, since the variable *Standard error* is more valuable for our analysis we keep it within the dataset. Further, in Figure A2 we detect a correlation of 0.95 between variables *Publication year* and *Avg data year*. For the purpose of the analysis, we decided to keep variable *Publication year*.

The rest of the matrix showcases a few higher correlations that could potentially influence the accuracy of our findings. Despite this, we keep the related variables and strive to control this correlation through BMA modifications and selecting suitable priors on the model space. Further, according to the Variance Inflation Factor (VIF), the rest of the variables are suitable for further examination. Overall, we execute BMA with a total of 37 explanatory variables.

Defining Prior Distributions Before delving into the BMA application, we outline prior distributions on individual regression parameters and model probabilities. Given our scant prior information about the parameter space, we decide, in line with Eicher et al. (2011), to adopt a widely used default prior - the unit information prior (UIP). However, in terms of our prior choice for the model space, we deviate from the conventional approach of using a uniform model. Instead, we opt for the collinearity-adjusted dilution model prior proposed by George et al. (2010). This prior accommodates the presence of correlation between explanatory variables, addressing multicollinearity by appropriately down-weighting the posterior probabilities of models featuring highly correlated covariates (Hasan et al., 2018).

With our chosen settings, BMA estimated using the unit information prior and dilution model prior serves as our baseline approach. However, we make one additional alteration to the baseline BMA. Considering the substantial imbalance in our dataset with regards to the number of estimates each study reports, we adhere to Havranek et al. (2018)'s of weighting our baseline model by the inverse of the number of estimates in each study.

Over and above the baseline model, we introduce several robustness checks. Drawing inspiration from Gechert et al. (2019), we compare our baseline model with two alternative settings - *Random & BRIC* and *Random & HQ*. For the estimation of the baseline BMA and its alternations, we employ an R package BMS developed by Zeugner and Feldkircher (2015).

4.2.2 Frequentist Model Averaging

Turning to Frequentist Model Averaging (FMA), the enduring statistical debate between bayesian and frequentist methodologies arises from their distinct strengths and limitations (Bayarri and Berger, 2004). Consequently, we see merit in leveraging both methods. Following the lead of Havranek et al. (2017), we choose FMA as complementary approach to BMA. The primary appeal of FMA is its reliance solely on data and the absence of a need for prior specification,

leading to a more objective approach. For its implementation, we refer to the code provided in the online appendix by Havranek et al. (2021).

4.2.3 Frequentist Check

Ultimately, we conduct a frequentist check by estimating Equation 8 using OLS with standard errors clustered at the study level. In this analysis, we solely incorporate variables with a prior inclusion probability of 0.5 or higher from the baseline BMA approach, indicating a minimum threshold of weak importance for these variables.

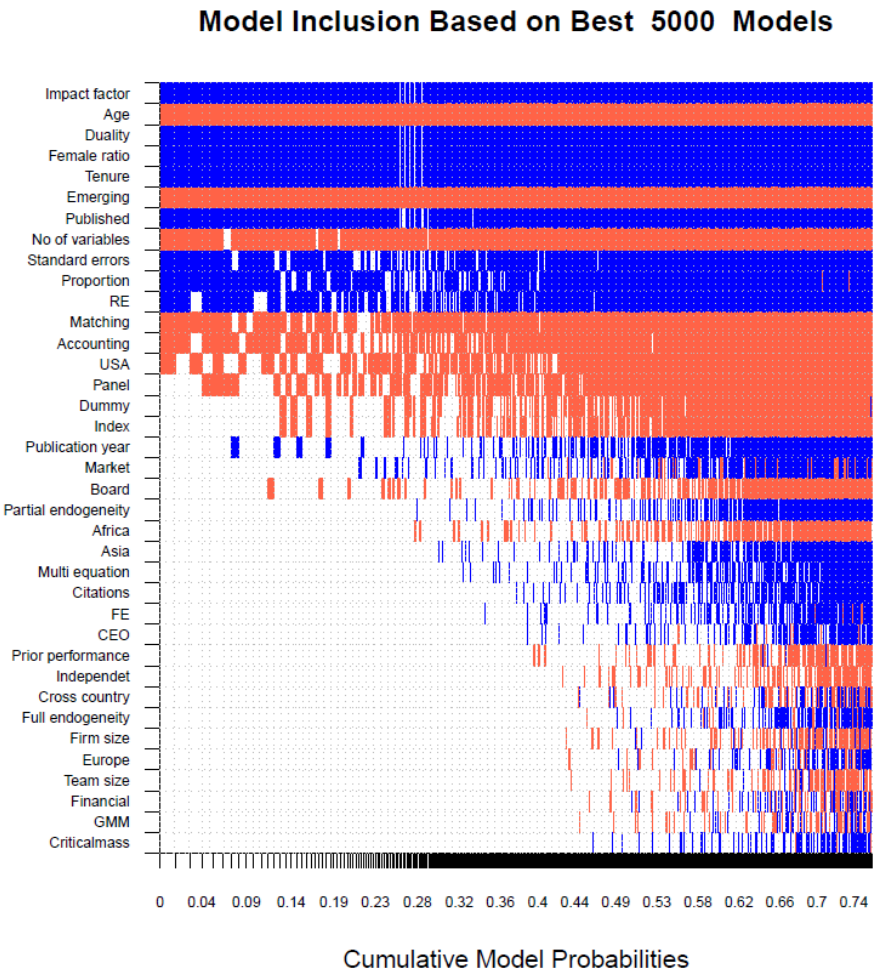
4.3 Results

The outcomes of our BMA estimation are visualised in Figure 1. The vertical axis sorts the independent variables based on their posterior inclusion probabilities, ranking them from highest to lowest. As a result, the top section of the graphic is occupied by the most impactful predictors. In contrast, the horizontal axis showcases the models with the highest efficacy. The breadth of each column is dictated by the posterior model probability. The graphical display follows a specific colour coding: blank spaces denote the exclusion of certain variables from the model, a red colour signifies a negative coefficient for a specific variable, while a blue colour implies a positive coefficient.

Thirteen variables emerge as significant contributors to the heterogeneity within the computed PCCs, each with a posterior inclusion probability exceeding 0.5. The BMA visualisation is reinforced by the quantitative information in Table 6, corroborating the insights derived from the plot. According to the classification system developed by Kass and Raftery (1995), variables *Female ratio*, *Impact factor*, *Tenure*, *Age*, and *Duality* carry decisive influence. Our findings also highlight a strong impact of variable *Emerging*. Variables representing *Standard error*, *Published*, and *No of variables* suggest a positive influence. Lastly, variables *RE*, *Matching*, *Proportion* and *Accounting* present a weak effect on the relationship under analysis.

Publication Bias and Data Characteristics Analysis in the Section 3 reveals the prevalence of a positive publication bias within the existing body of literature examining the financial performance implications of female leadership. The bias retains its significance and direction, indicated by a PMP and a PIP corresponding to the *Standard error* variable, as our analysis

Figure 5: Visualisation of BMA estimation



Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability.

extends to encompass more explanatory variables. Notwithstanding, these additional variables seem to modulate the overall intensity of the publication bias, implying that its effect, though still present, is softened compared to our initial observations. The result is supported by the FMA estimation, which recognises the significance at the 5% level and by the frequentist check where *Standard error* is significant at the 1% level.

Continuing with data characteristics, the number of variables included in the initial analysis influences the results. Studies implementing higher numbers of variables within their models yield more negative results.

Variable *Panel* has PIP of 0.464 hence the capitalisation of panel data does not contribute to unfolding disparities in the estimated PCCs.

Publication Characteristics Our analysis reveals scant evidence to suggest that the number of citations influences the reported effect estimates. The lack of a conclusive impact of the number of citations on the heterogeneity of studies might be attributed to the unique characteristics of the most cited papers within our sample. Frequently, these highly referenced papers are known not for presenting substantially different estimates but rather for introducing novel estimation approaches that have become industry standards over the years.

Similarly the publication year of the studies does not introduce heterogeneity in our analysis. However, salient factors contributing to the heterogeneity of PCCs are the ratio of female authors of the original study, its published status, and the quality of the journal (measured by the Journal Citation Reports impact factor), as indicated by the high posterior inclusion probability value associated with these variables. All of them have a positive impact on the studied relationship.

When the ratio of female authors in the study increases in proportion to male authors, studies tend to exhibit more positive results. We suggest several plausible explanations for this observation: (i) female authors may contribute unique insights into their research, informed by their personal experiences or comprehension of gender-related concerns; (ii) consistent with research indicating an inherent bias toward those sharing similar characteristics, female authors might unintentionally highlight the positive aspects of female leadership; (iii) papers authored by females, particularly those demonstrating positive implications of female leadership, might be more likely to secure publication due to prevailing societal interest or academic interest in such findings. Unfortunately, our analysis is not able to conclusively support or refute any of these explanations, leaving them all equally possible.

Further, studies published in reviewed journals tend to produce more positive effects than those that remain unpublished. The publication process typically involves a peer review, which generally guarantees a certain degree of methodological precision. Conversely, unpublished studies, such as dissertations or working papers, may not have been subjected to this stringent review process. This difference could introduce heterogeneity in the data, leading to variations in the results of the studies.

According to the BMA, the impact factor has a decisive positive effect on the studied relationship. This finding is also endorsed by FMA and frequentist check. The positive sign of

PMP and the RePEc factor rising with the impact of a journal posit that studies published in higher-tier journals are likely to report higher estimates of the effect female leaders have on the company's financial outcomes.

Estimation Methods Based on our findings, it is evident that estimation methods generally yield similar effect estimates, with the exception of the random estimation approach, which shows noticeable variation. RE subtly influences the estimated relationship and generates more positive results than the other methods. While it may seem unexpected given the varying capabilities of these methods to address endogeneity (Adams et al., 2009). Our result aligns with the conclusions drawn by Wang et al. (2019) or Lafuente and Vaillant (2019), who suggest that even though there are differences in effect magnitudes and directions between estimation techniques, they are generally not statistically significant.

A plausible rationale for the minimal influence of various estimation methods could be linked to the inherent challenges in accurately utilising more advanced techniques, such as the instrumental variable approach. These sophisticated methods demand a comprehensive understanding and careful application, which may contribute to their limited impact on the study outcomes (Peni, 2014; Larcker and Rusticus, 2010).

Moreover, investigations that utilise matching methods or examine the relationship within a matched sample tend to produce more negative outcomes compared to those that do not employ such techniques.

Analytical Design Accounting for endogeneity in studies does not significantly shift the estimated partial correlation coefficients, a finding that resonates with the outcomes derived from applying various estimation methods. This observation aligns with the conclusions of Liu et al. (2014), who utilise varying degrees of endogeneity control and noted effects that were either non-significant or exhibited similar direction and magnitude.

The geographical origin of the data used in the analysis does not seem to play an essential role in introducing heterogeneity in the results, as none of the variables exhibit significance.

Analyses focused on emerging markets yield more negative results compared to non-emerging markets. This finding aligns with the existing literature, which suggests that in countries with underdeveloped infrastructure, female representation in the upper echelon faces greater obsta-

cles, resulting in less favourable outcomes than their male counterparts (Ararat and Yurtoglu, 2021).

Surprisingly, distinguishing between financial and non-financial industries in primary studies does not introduce notable heterogeneity in the data. This finding challenges the notion that financial and non-financial sectors should be analysed separately.

Table 6: Results of BMA, FMA and frequentist check estimations

	Bayesian Model Averaging			**Frequentist Model Averaging**			**Frequentist Check**		
	Post. Mean	*Post. SD*	*PIP*	*Coef.*	*SD*	*p-value*	*Coef.*	*SD*	*p-value*
Standard error	0.3156	0.1905	0.8015	0.2772	0.1332	0.037	0.4748	0.1376	0.0006
Data Characteristics									
No of variables	-0.0068	0.003	0.9025	-0.0055	0.0022	0.012	-0.0043	0.0012	0.0005
Panel	-0.0109	0.0131	0.4644	0.0049	0.0096	0.610			
Publication characteristics									
Publication year	0.0017	0.0041	0.174	0.0137	0.0040	0.001			
Published	0.0436	0.0171	0.9363	0.0604	0.0181	0.001	0.0421	0.0113	0.0002
Impact factor	0.1337	0.0232	1.0000	0.0666	0.0261	0.011	0.0809	0.0171	0.0000
Citations	0.0001	0.001	0.0358	-0.0023	0.0032	0.472			
Female ratio	0.0347	0.0079	0.9996	-0.0034	0.0092	0.712	0.019	0.0073	0.0094
Estimation Method									
FE	0.0002	0.0021	0.0235	-0.0142	0.0119	0.233			
RE	0.0279	0.0199	0.7321	0.0156	0.0153	0.308	0.0278	0.0105	0.0084
Multi equation	0.0007	0.0038	0.0406	0.0048	0.0163	0.768			
GMM	-0.0000	0.0008	0.0092	-0.0082	0.0157	0.601			
Matching	-0.0449	0.0328	0.7177	-0.0163	0.0130	0.210	-0.0288	0.0209	0.1688
Analytical Design									
Partial endogeneity	0.0007	0.0034	0.0597	0.0142	0.0118	0.229			
Full endogeneity	0.0001	0.0012	0.0124	0.0193	0.0145	0.183			
Cross country	-0.0000	0.0011	0.0128	0.0228	0.0137	0.096			
USA	-0.0131	0.0149	0.4905	0.0119	0.0163	0.465			
Europe	0.0000	0.0009	0.0107	0.0039	0.0144	0.787			
Asia	0.0007	0.0036	0.0472	0.0258	0.0113	0.022			
Africa	-0.0016	0.0072	0.0585	-0.0111	0.0168	0.509			
Emerging	-0.0294	0.0085	0.9867	-0.0271	0.0124	0.029	-0.0266	0.0053	0.0000
Financial	0.0000	0.0011	0.0094	0.0203	0.0123	0.099			
Variable Specifications									
CEO	0.0002	0.0019	0.0174	-0.0040	0.0106	0.706			
Board	-0.0017	0.0054	0.1122	0.0095	0.0082	0.247			
Dummy	-0.0049	0.0096	0.2344	-0.0299	0.0127	0.019			
Index	-0.0074	0.0152	0.2183	-0.0478	0.0161	0.003			
Proportion	0.0162	0.0107	0.7478	-0.0290	0.0117	0.013	-0.0024	0.0051	0.6393
Criticalmass	0.0000	0.0013	0.0089	0.0045	0.0093	0.628			
Accounting	-0.0108	0.0097	0.6175	-0.0167	0.0140	0.233	-0.0255	0.0056	0.0000
Market	0.0015	0.0053	0.1202	0.0057	0.0146	0.696			
Control Variables									
Firm size	-0.0000	0.0013	0.0122	-0.0338	0.0091	0.000			
Team size	-0.0000	0.0007	0.0097	-0.0178	0.0078	0.022			
Independent	-0.0000	0.001	0.0141	-0.0080	0.0077	0.299			
Duality	0.0463	0.0064	1.0000	0.0345	0.0075	0.000	0.0229	0.0058	0.0000
Age	-0.0684	0.0105	1.0000	-0.0455	0.0106	0.000	-0.0425	0.0103	0.0000
Tenure	0.0495	0.0113	0.9989	0.0129	0.0109	0.237	0.0149	0.0115	0.1942
Prior performance	-0.0000	0.0012	0.0152	0.0037	0.0098	0.706			
Studies	96	96	96	96	96	96	96	96	96
Observations	1,131	1,131	1,131	1,131	1,131	1,131	1,131	1,131	1,131

Notes: Post. mean = posterior mean, Post. SD = posterior standard deviation, PIP = posterior inclusion probability. The posterior mean in Bayesian model averaging (or alternatively the estimated coefficient in frequentist model averaging and frequentist check) denotes the marginal effect of a study characteristic on the estimate of beta reported in the literature. For detailed description of all the variables see Table 5.

Variable Specifications Studies focusing on the proportion of females in top leadership positions yield more positive findings than those examining other factors, such as the number or presence of females or gender diversity within the upper echelon. In contrast, including dummy variables or indices related to gender does not seem to introduce significant heterogeneity in the estimates. These findings suggest that the mere presence of female leaders or gender diversity alone may not significantly impact outcomes. Instead, the relative proportion of female leaders in relation to the overall leadership is vital.

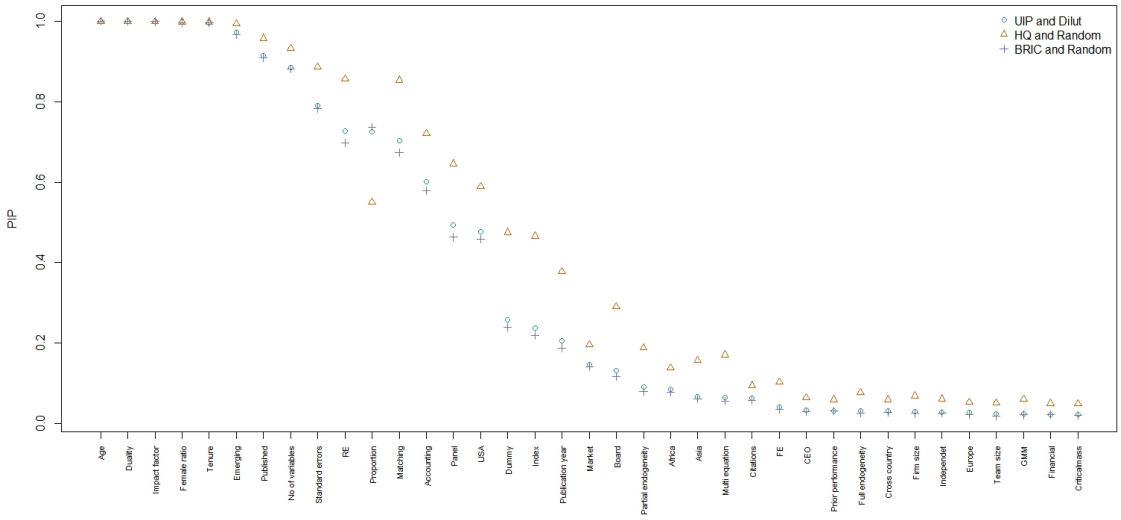
This discovery underscores the significance of identifying an appropriate threshold for female representation to elicit meaningful change, thereby supporting the critical mass theory. To capture the concept, we introduced a variable *Criticalmass* for recognising studies that factor this threshold into their analyses. However, this variable did not exhibit any discernible influence within our dataset. We propose the following explanation for this observation. Most primary studies incorporating critical mass in analysis focus on companies globally or across diverse industries. The variations in team sizes within these contexts may significantly differ, and with that, the threshold at which the impact of female representation becomes observable.

The variation in leadership positions within companies does not significantly affect the estimated PCCs. In many companies, leadership roles such as CEO, board member or top manager are often interchangeable, with many shared responsibilities and powers. As a result, specific categorisation may not have a significant differential impact on the outcome. The perfect example is a CEO duality, where the chief executive officer also serves as a board member. A similar conclusion is reached in the study by Peni (2014) or Ullah et al. (2020), where analysing the alternative leadership position did not change its effect on the studied relationship.

Control Variables Among the multitude of explanatory variables examined in the primary studies, controlling for factors such as firm size, director independence, team size and prior performance does not introduce heterogeneity in the results. However, three variables emerge as influential contributors to heterogeneity: tenure, age, and CEO duality. Notably, the analysis tends to produce more positive outcomes when the CEO also serves as a board member. The inclusion of age in the study often results in more negative outcomes. In contrast, when tenure, or the length of service, is factored into the analysis, the outcomes tend to be more positive.

As previously mentioned, we also performed robustness checks by implementing BMA with different priors, thereby enhancing the dependability of our results. Figure 6 provides a visual representation of the posterior inclusion probabilities of individual variables for all three of these models.

Figure 6: BMA estimation with different priors



The findings from the Frequentist Model Averaging closely parallel those obtained from the Bayesian Model Averaging method. The FMA results display the same direction and significance of effects for almost all variables under scrutiny. Likewise, the results of the frequentist check correspond with those from the BMA estimation, with a few exceptions. Specifically, variables *matching*, *proportion*, and *tenure* are considered statistically insignificant in the frequentist analysis.

5 Best Practice Estimate

Following a thorough assessment of publication bias and heterogeneity within the collected estimates, our objective is to generate the best practice estimate. Viewed as the final takeaway from the meta-analysis, the best practice estimate represents a comprehensive distillation of our findings. In essence, we derive it through the execution of a linear regression where each incorporated variable is assigned a preferred value. This selected value corresponds to the

sample minimum for an undesirable practice, the sample maximum for the best practice, and the sample mean for scenarios without any preference.

However, it is crucial to acknowledge that this endeavour is experimental. Therefore, we would like to highlight the inherent limitations of this method at the onset. Firstly, delineating a best practice estimate carries an element of subjectivity as we arbitrate over the facets of an ideal study. Secondly, the best practice estimate in our context lacks explicit economic significance, given the application of PCCs.

In our assessment, an optimal estimation should fulfill certain criteria. To counter publication bias, the standard error is set to its sample minimum. Considering publication impact, we believe that studies from journals with high JCR impact factors are likely to provide the best estimates. Hence, we set publication traits at their sample maximum, except for the number of citations, which we set to the sample mean, as high citations may not always indicate superior quality. To address potential endogeneity bias, we prioritise panel data over cross-sectional data and employ estimation methods like GMM and multi-equation approach. For these variables, we use the sample maximum. Regarding financial measures and female leadership positions, we remain neutral and set these variables to the dataset mean. We advocate measuring female leadership as a proportion, directly representing the ratio of female leaders. All dummies for control variables are assigned a value of 1, even though not all of them reported significance in the BMA excercise we consider them important to build model. For the remaining variables, we maintain their sample mean without any preference.

Table 7: Best practice estimates implied for different data sets

	Predicted Estimate	95% Confidence Interval
General	0.1168	[0.051;0.183]
Board	0.1044	[0.044;0.164]
CEO	0.1199	[0.06;0.179]
TMT	0.1312	[0.065;0.196]

Notes: The table presents predicted estimates for different leadership positions. The 95% confidence intervals in square brackets are approximate and constructed using the standard errors estimated by OLS with standard errors clustered at the study level.

According to our stated preferences, Table 7 presents multiple best practice estimates along with their respective 95% confidence intervals. The computed partial correlation coefficient value is 0.1168, and results of studies focusing on gender diversity in leadership positions align

closely with this optimal estimate. All best practice estimates report stronger effect compared to previous chapters. Even though the effect is stronger, it is still minimal. After closer examination, we find the main positive drivers of the effect to be *Impact factor*, *Published*, *Duality*, *Tenure* and negative one *Age*. These results suggest that once we control for such factors with caution the final impact of female leaders on the financial performance of companies is positive and non-zero.

6 Concluding Remarks

The objective of this meta-analysis was to examine the effect female leaders have on the financial outcomes of the company. While we acknowledged the existence of four previous meta-analyses on this subject they did not address it in substantial depth. In our study, we thoroughly evaluated the subject by collecting 1,131 partial correlation coefficients from 96 unique studies on the up-to-date dataset.

In the first part of the study we addressed the concern of publication bias (Stanley 2005) by employing a visual assessment tool called Funnel Plot along with a variety of statistical tests, including the FAT-PAT, non-linear techniques, and methods allowing for endogeneity. As a result, only some of the tests yielded significant estimates of publication bias spanning from 0.2697 to 0.3313. Nevertheless, these significant estimates implied only small evidence of publication bias. In addition, these methods also provided an estimate of the effect beyond bias. Most of the approaches we implemented yielded statistically significant and positive estimates, ranging from 0.0098 to 0.0220. Given that we were dealing with PCCs, this effect was considered negligible. In addition, we calculated best practice estimates indicating a stronger final effect but still close to zero.

In the second part of our study, we took into account factors other than publication bias that could potentially influence the variability in the estimated PCCs derived from primary studies. We employed Bayesian Model Averaging weighted by inverse number of estimates as our baseline model, and controlled for 37 variables. Out of these, thirteen were identified as significant moderators of the relationship under study. This analytical process confirmed the presence of positive publication bias. However, its impact was not that substantial when we accounted for additional characteristics. Factors that exerted a positive influence included the

ratio of female authors, the impact factor of the journal, and the proportion of female leadership, along with the published status of the article, implementation of random estimation, duality of CEOs or including personal trait - tenure. On the other hand, certain factors decreased the effect. These included the number of variables in the primary analysis, a focus on emerging economies, the use of matched samples or accounting-based financial measures, and the inclusion of the leaders' age in the analysis.

Apart from their academic importance, our findings also bear practical relevance. They offer valuable insights for policy makers, specifically in the areas of gender diversity and inclusion, leadership progression, and succession planning.

The primary constraint of our study stems from the use of partial correlation coefficients to achieve comparable effect sizes, a common practice in meta-analyses, which presents challenges for interpretation. In light of this critique, our preference would be to utilise original effect estimates. However, primary studies investigating gender diversity in senior leadership and its impact on corporate financial outcomes employ a variety of measures for both dependent and independent variables. Further complicating the situation are incomplete or inconsistent definitions of variables included in these studies. This lack of standardisation and clarity can introduce additional sources of bias making it more difficult to draw accurate and reliable conclusions.

Future research in this field could greatly benefit from the standardisation of measurement techniques used to evaluate the impact of female leadership on financial performance. This would enhance comparability across studies, strengthening the validity of meta-analytical findings. And it would not only improve the replicability of studies but also deepen the understanding of the factors influencing the relationship between female leadership and financial performance.

References

Abdelzaher, A. and Abdelzaher, D. (2019). Women on boards and firm performance in egypt: post the arab spring. *The Journal of Developing Areas*, 53(1).

Abdullah, S. N. and Ismail, K. N. I. K. (2016). Women directors, family ownership and earnings management in malaysia. *Asian Review of Accounting*, 24(4):525–550.

Adams, R. B. (2016). Women on boards: The superheroes of tomorrow? *The Leadership Quarterly*, 27(3):371–386.

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, S. M., Gupta, A., and Leeth, J. D. (2009). Are female executives over-represented in precarious leadership positions? *British Journal of Management*, 20(1):1–12.

Ahern, K. R. and Dittmar, A. K. (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. *The quarterly journal of economics*, 127(1):137–197.

Ahmad, M., Raja Kamaruzaman, R. N. S., Hamdan, H., and Annuar, H. A. (2020). Women directors and firm performance: Malaysian evidence post policy announcement. *Journal of Economic and Administrative Sciences*, 36(2):97–110.

Ahmadi, A., Nakaa, N., and Bouri, A. (2018). Chief executive officer attributes, board structures, gender diversity and firm performance among french cac 40 listed firms. *Research in International Business and Finance*, 44:218–226.

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

Ararat, M. and Yurtoglu, B. B. (2021). Female directors, board committees, and firm performance: Time-series evidence from turkey. *Emerging Markets Review*, 48:100768.

Atif, M. et al. (2021). Does board gender diversity affect renewable energy consumption? *Journal of Corporate Finance*, 66.

Bajerova, V. (2021). Female leadership and financial performance: Evidence from the czech republic.

Baselga-Pascual, L. and Vahamaa, E. (2021). Female leadership and bank performance in latin america. *Emerging Markets Review*, 48:100807.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80.

Belaounia, S., Tao, R., and Zhao, H. (2020). Gender equality's impact on female directorsâ€™ efficacy: A multi-country study. *International Business Review*, 29(5):101737.

Bennouri, M., Chtioui, T., Nagati, H., and Nekhili, M. (2018). Female board directorship and firm performance: What really matters? *Journal of Banking & Finance*, 88:267–291.

Blau, F. D. and Kahn, L. M. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of labor Economics*, 15(1):1–42.

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

Bonn, I., Yoshikawa, T., and Phan, P. H. (2004). Effects of board structure on firm performance: A comparison between japan and australia. *Asian Business & Management*, 3:105–125.

Brahma, S., Nwafor, C., and Boateng, A. (2021). Board gender diversity and firm performance: The uk evidence. *International Journal of Finance & Economics*, 26(4):5704–5719.

Campbell, K. and Minguez-Vera, A. (2008). Gender diversity in the boardroom and firm financial performance. *Journal of Business Ethics*, 83:435–451.

Campbell, K. and Minguez Vera, A. (2010). Female board appointments and firm valuation: Short and long-term effects. *Journal of Management & Governance*, 14:37–59.

Cannella Jr, A. A., Park, J.-H., and Lee, H.-U. (2008). Top management team functional background diversity and firm performance: Examining the roles of team member colocation and environmental uncertainty. *Academy of management Journal*, 51(4):768–784.

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

Carter, D. A., D'Souza, F., Simkins, B. J., and Simpson, W. G. (2010). The gender and ethnic diversity of us boards and board committees and firm financial performance. *Corporate Governance: An International Review*, 18(5):396–414.

Carter, D. A., Simkins, B. J., and Simpson, W. G. (2003). Corporate governance, board diversity, and firm value. *Financial review*, 38(1):33–53.

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

Chadwick, I. C. and Dawson, A. (2018). Women leaders and firm performance in family businesses: An examination of financial and nonfinancial outcomes. *Journal of family business strategy*, 9(4):238–249.

Chijoke-Mgbame, A. M., Boateng, A., and Mgbame, C. O. (2020). Board gender diversity, audit committee and financial performance: evidence from nigeria. *Accounting Forum*, 44(3).

Darmadi, S. (2011). Board diversity and firm performance: The indonesian evidence. *Corporate ownership and control Journal*, 8.

Darmadi, S. (2013). Do women in top management affect firm performance? evidence from indonesia. *Corporate Governance: The international journal of business in society*, 13(3):288–304.

Dezsö, C. L. and Ross, D. G. (2012). Does female representation in top management improve firm performance? a panel data investigation. *Strategic management journal*, 33(9):1072–1089.

Doucouliagos, C. (2011). How large is large? preliminary and relative guidelines for interpreting partial correlations in economics.

Doucouliagos, C. and Laroche, P. (2003). What do unions do to productivity? a meta-analysis. *Industrial Relations: A Journal of Economy and Society*, 42(4):650–691.

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.

Duppati, G., Rao, N. V., Matlani, N., Scrimgeour, F., and Patnaik, D. (2020). Gender diversity and firm performance: evidence from india and singapore. *Applied Economics*, 52(14):1553–1565.

Eagly, A. H. and Carli, L. L. (2003). The female leadership advantage: An evaluation of the evidence. *The leadership quarterly*, 14(6):807–834.

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.

Ellwood, S. and Garcia-Lacalle, J. (2018). New development: Women with altitudeâ€"exploring the influence of female presence and leadership on boards of directors. *Public Money & Management*, 38(1):73–78.

Erkut, S., Kramer, V. W., and Konrad, A. M. (2008). 18. critical mass: Does the number of women on a corporate board make a difference. *Women on corporate boards of directors: International research and practice*, 222.

Fauzi, F. and Locke, S. (2012). Board structure, ownership structure and firm performance: A study of new zealand listed-firms.

Fernando, G. D., Jain, S. S., and Tripathy, A. (2020). This cloud has a silver lining: Gender diversity, managerial ability, and firm performance. *Journal of business research*, 117:484–496.

Flabbi, L., Piras, C., and Abrahams, S. (2017). Female corporate leadership in latin america and the caribbean region: Representation and firm-level outcomes. *International Journal of Manpower*, 38(6):790–818.

Furukawa, C. (2019). Publication bias under aggregation frictions: Theory, evidence, and a new correction method. *EconStor Preprints*, 194798.

Galbreath, J. (2011). Are there gender-related influences on corporate sustainability? a study of women on boards of directors. *Journal of management & organization*, 17(1):17–38.

Galbreath, J. (2018). Is board gender diversity linked to financial performance? the mediating mechanism of csr. *Business & Society*, 57(5):863–889.

García-Meca, E., García-Sánchez, I.-M., and Martínez-Ferrero, J. (2015). Board diversity and its effects on bank performance: An international analysis. *Journal of banking & Finance*, 53:202–214.

García-Sánchez, I.-M., Gallego-Álvarez, I., and Zafra-Gómez, J.-L. (2021). Do independent, female and specialist directors promote eco-innovation and eco-design in agri-food firms? *Business Strategy and the Environment*, 30(2):1136–1152.

García-Sánchez, I.-M. and Martínez-Ferrero, J. (2019). Chief executive officer ability, corporate social responsibility, and financial performance: The moderating role of the environment. *Business Strategy and the Environment*, 28(4):542–555.

Gechert, S., HavrÅˇnek, T., HavrÅˇnkovÅˇ, Z., and Kolcunova, D. (2019). Death to the cobb-douglas production function. Technical Report 201, IMK Working Paper.

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

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

Gerber, A. and Malhotra, N. (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.

Girón, A., Kazemikhasragh, A., Cicchiello, A. F., and Panetti, E. (2021). Sustainability reporting and firmsâ€™ economic performance: Evidence from asia and africa. *Journal of the Knowledge Economy*, 12:1741–1759.

González, M., Guzmán, A., Pablo, E., and Trujillo, M. A. (2020). Does gender really matter in the boardroom? evidence from closely held family firms. *Review of Managerial Science*, 14:221–267.

Gregory-Smith, I., Main, B. G., and O'Reilly III, C. A. (2014). Appointments, pay and performance in uk boardrooms by gender. *The Economic Journal*, 124(574):F109–F128.

Gull, A. A., Atif, M., and Hussain, N. (2023). Board gender composition and waste management: Cross-country evidence. *The British Accounting Review*, 55(1):101097.

Hakimah, Y., Pratama, I., Fitri, H., Ganatri, M., and Sulbahrie, R. A. (2019). Impact of intrinsic corporate governance on financial performance of indonesian smes. *International Journal of Innovation, Creativity and Change Vol*, 7(1):32–51.

Harrison, D. A. and Klein, K. J. (2007). What's the difference? diversity constructs as separation, variety, or disparity in organizations. *Academy of management review*, 32(4):1199–1228.

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.

Haslam, S. A. et al. (2010). Investing with prejudice: The relationship between women's presence on company boards and objective and subjective measures of company performance. *British Journal of Management*, 21(2):484–497.

Havranek, T., Irsova, Z., Laslopova, L., and Zeynalova, O. (2020). The elasticity of substitution between skilled and unskilled labor: A meta-analysis. MPRA Paper 102598, University Library of Munich, Germany.

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

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

HavrĂˇnek, T., Irsova, Z., and Vlach, T. (2018). Measuring the income elasticity of water demand: The importance of publication and endogeneity biases. *Land Economics*, 94(2):259–283.

He, J. and Huang, Z. (2011). Board informal hierarchy and firm financial performance: Exploring a tacit structure guiding boardroom interactions. *Academy of Management Journal*, 54(6):1119–1139.

Hedija, V. and Němec, D. (2021). Gender diversity in leadership and firm performance: Evidence from the czech republic. *Journal of Business Economics and Management*, 22(1):156–180.

Hoobler, J. M., Masterson, C. R., Nkomo, S. M., and Michel, E. J. (2018). The business case for women leaders: Meta-analysis, research critique, and path forward. *Journal of management*, 44(6):2473–2499.

Hopp, C., Wentzel, D., and Rose, S. (2020). Chief executive officers' appearance predicts company performance, or does it? a replication study and extension focusing on ceo successions. *The Leadership Quarterly*, page 101437.

Horváth, R., Spirollari, P., et al. (2012). Do the board of directorsâ€™ characteristics influence firmâ€™s performance? the us evidence. *Prague economic papers*, 4(2):470–486.

Investopedia (2023a). Emerging market economy.

Investopedia (2023b). Financial performance definition.

Investopedia (2023c). Q ratio or tobin's q: Definition, formula, uses, and examples.

Investopedia (2023d). Return on equity (roe) vs. return on assets (roa): What's the difference?

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

Isidro, H. and Sobral, M. (2015). The effects of women on corporate boards on firm value, financial performance, and ethical and social compliance. *Journal of Business Ethics*, 132:1–19.

Joecks, J., Pull, K., and Vetter, K. (2013). Gender diversity in the boardroom and firm performance: What exactly constitutes a â€ścritical mass?â€ť. *Journal of business ethics*, 118:61–72.

Julizaerma, M. K. and Sori, Z. M. (2012). Gender diversity in the boardroom and firm performance of malaysian public listed companies. *Procedia-Social and Behavioral Sciences*, 65:1077–1085.

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

Koop, G. (2003). *Bayesian Econometrics*. Wiley.

Krishnan, H. A. and Park, D. (2005). A few good womenâ€"on top management teams. *Journal of business research*, 58(12):1712–1720.

Kweh, Q. L., Ahmad, N., Ting, I. W. K., Zhang, C., and Hassan, H. (2019). Board gender diversity, board independence and firm performance in malaysia. *Institutions and Economies*, pages 1–20.

Lafuente, E. and Vaillant, Y. (2019). Balance rather than critical mass or tokenism: Gender diversity, leadership and performance in financial firms. *International Journal of Manpower*, 40(5):894–916.

Lantz, B., Bredehorst-Carlsson, P., and Johansson, J. (2012). Incentive schemes and female leadership in financial firms. *Lantz, B., Bredehorst-Carlsson, P. & Johansson, J.(2013).â€śIncentive Schemes and Female Leadership in Financial Firmsâ€ť, Corporate Board: Role, Duties and Composition*, 9:40–49.

Larcker, D. F. and Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics*, 49(3):186–205.

Lee, P. M. and James, E. H. (2007). She'-e-os: gender effects and investor reactions to the announcements of top executive appointments. *Strategic Management Journal*, 28(3):227–241.

Lindenberg, E. B. and Ross, S. A. (1981). Tobin's q ratio and industrial organization. *Journal of Business*, pages 1–32.

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

Liu, Y., Wei, Z., and Xie, F. (2014). Do women directors improve firm performance in china? *Journal of corporate finance*, 28:169–184.

Low, Daniel CM, H. R. and Whiting, R. H. (2015). Board gender diversity and firm performance: Empirical evidence from hong kong, south korea, malaysia and singapore. *Pacific-Basin Finance Journal*, 35:381–401.

Lückerath-Rovers, M. (2013). Women on boards and firm performance. *Journal of Management & Governance*, 17:491–509.

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

Marimuthu, M. and Kolandaisamy, I. (2009). Can demographic diversity in top management team contribute for greater financial performance an empirical discussion.

Martínez, M. d. C. V. and Rambaud, S. C. (2019). Women on corporate boards and firm's financial performance. In *Women's Studies International Forum*, volume 76, page 102251. Elsevier.

Matousek, J., Havranek, T., and Irsova, Z. (2019). Individual discount rates: A meta-analysis of experimental evidence. EconStor Preprints 194617, ZBW - Leibniz Information Centre for Economics.

Medenbach, O. et al. (2001). Refractive index and optical dispersion of rare earth oxides using a small-prism technique. *Journal of Optics A: Pure and Applied Optics*, 3(3):174.

Mersland, R. and Strøm, R. Ø. (2009). Performance and governance in microfinance institutions. *Journal of Banking & Finance*, 33(4):662–669.

Miller, T. and Triana, M. d. C. (2009). Demographic diversity in the boardroom: Mediators of the board diversityâ€“firm performance relationship. *Journal of Management studies*, 46(5):755–786.

Moral-Benito, E. (2012). Determinants of economic growth: a bayesian panel data approach. *Review of Economics and Statistics*, 94(2):566–579.

Nadeem, M., Gyapong, E., and Ahmed, A. (2020). Board gender diversity and environmental, social, and economic value creation: Does family ownership matter? *Business Strategy and the Environment*, 29(3):1268–1284.

Naseem, M. A., Lin, J., Rehman, R. u., Ahmad, M. I., and Ali, R. (2019). Does capital structure mediate the link between ceo characteristics and firm performance? *Management Decision*, 58(1):164–181.

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.

Nekhili, M., Gull, A. A., Nagati, H., and Chtioui, T. (2018). Beyond gender diversity: How specific attributes of female directors affect earnings management. *The British Accounting Review*, 50(3):255–274.

Nguyen, T., Locke, S., and Reddy, K. (2015). Does boardroom gender diversity matter? evidence from a transitional economy. *International Review of Economics & Finance*, 37:184–202.

Overveld, M. N. (2012). Board diversity and financial firm performance in dutch listed firms. Master's thesis, University of Twente.

Papangkorn, S. et al. (2021). Female directors and firm performance: Evidence from the great recession. *International Review of Finance*, 21(2):598–610.

Pathan, S. and Faff, R. (2013). Does board structure in banks really affect their performance? *Journal of Banking Finance*, 37(5):1573–1589.

Peni, E. (2014). Ceo and chairperson characteristics and firm performance. *Journal of Management & Governance*, 18:185–205.

Perryman, A. A., Fernando, G. D., and Tripathy, A. (2016). Do gender differences persist? an examination of gender diversity on firm performance, risk, and executive compensation. *Journal of Business Research*, 69(2):579–586.

Pletzer, J. L., Nikolova, R., Kedzior, K. K., and Voelpel, S. C. (2015). Does gender matter? female representation on corporate boards and firm financial performance-a meta-analysis. *PloS one*, 10(6):e0130005.

Post, C. and Byron, K. (2015). Women on boards and firm financial performance: A meta-analysis. *Academy of management Journal*, 58(5):1546–1571.

Pucheta-Martínez, M. C. and Gallego-Álvarez, I. (2020). Do board characteristics drive firm performance? an international perspective. *Review of Managerial Science*, 14(6):1251–1297.

Qian, M. (2016). Women's leadership and corporate performance. *Asian Development Bank Economics Working Paper Series*, (472).

Reddy, S. and Jadhav, A. M. (2019). Gender diversity in boardrooms–a literature review. *Cogent Economics & Finance*, 7(1):1644703.

Reguera-Alvarado, N., De Fuentes, P., and Laffarga, J. (2017). Does board gender diversity influence financial performance? evidence from spain. *Journal of business ethics*, 141:337–350.

Ren, T. and Wang, Z. (2011). Female participation in tmt and firm performance: evidence from chinese private enterprises. *Nankai Business Review International*, 2(2):140–157.

Richard, O. C., Barnett, T., Dwyer, S., and Chadwick, K. (2004). Cultural diversity in management, firm performance, and the moderating role of entrepreneurial orientation dimensions. *Academy of management journal*, 47(2):255–266.

Robb, A. M. and Watson, J. (2012). Gender differences in firm performance: Evidence from new ventures in the united states. *Journal of Business Venturing*, 27(5):544–558.

Rose, C. (2007). Does female board representation influence firm performance? the danish evidence. *Corporate governance: An international review*, 15(2):404–413.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Shrader, C. B., Blackburn, V. B., and Iles, P. (1997). Women in management and firm financial performance: An exploratory study. *Journal of managerial issues*, pages 355–372.

Shukeri, S. N., Shin, O. W., and Shaari, M. S. (2012). Does board of director's characteristics affect firm performance? evidence from malaysian public listed companies. *International business research*, 5(9):120.

Siegel, J. I., Kodama, N., and Halaburda, H. (2014). The unfairness trap: A key missing factor in the economic theory of discrimination. Working Paper 13-082, Harvard Business School Strategy Unit.

Simionescu, L. N., Gherghina, Ĺ. C., Tawil, H., and Sheikha, Z. (2021). Does board gender diversity affect firm performance? empirical evidence from standard & poorâ€™s 500 information technology sector. *Financial Innovation*, 7(1):1–45.

Solal, I. and Snellman, K. (2019). Women donâ€™t mean business? gender penalty in board composition. *Organization Science*, 30(6):1270–1288.

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

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

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

Stanley, T. D. and Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. routledge.

Stanley, T. D. and Doucouliagos, H. (2017). Neither fixed nor random: weighted least squares meta-regression. *Research synthesis methods*, 8(1):19–42.

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.

Steffensmeier, D. J., Schwartz, J., and Roche, M. (2013). Gender and twenty-first-century corporate crime: Female involvement and the gender gap in enron-era corporate frauds. *American Sociological Review*, 78(3):448–476.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association*, 54(285):30–34.

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

Sterne, J. A. C., Becker, B. J., and Egger, M. (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, chapter The.

Strøm, R. Ø., D'Espallier, B., and Mersland, R. (2014). Female leadership, performance, and governance in microfinance institutions. *Journal of Banking & Finance*, 42:60–75.

Terjesen, S., Couto, E. B., and Francisco, P. M. (2016). Does the presence of independent and female directors impact firm performance? a multi-country study of board diversity. *Journal of Management & Governance*, 20:447–483.

Trinh, V. Q. et al. (2018). Female leadership and value creation: Evidence from london stock exchange. *SSRN*.

Ujunwa, A. (2012). Board characteristics and the financial performance of nigerian quoted firms. *Corporate Governance: The international journal of business in society*, 12(5):656–674.

Ullah, I., Fang, H., and Jebran, K. (2020). Do gender diversity and ceo gender enhance firmâ€™s value? evidence from an emerging economy. *Corporate Governance: The International Journal of Business in Society*, 20(1):44–66.

Unite, A. A., Sullivan, M. J., and Shi, A. A. (2019). Board diversity and performance of philippine firms: Do women matter? *International Advances in Economic Research*, 25:65–78.

Uyar, A. et al. (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, C. 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.

Vairavan, A. and Zhang, G. P. (2020). Does a diverse board matter? a mediation analysis of board racial diversity and firm performance. *Corporate Governance: The international journal of business in society*, 20(7):1223–1241.

Vu, T.-H., Nguyen, V.-D., Ho, M.-T., and Vuong, Q.-H. (2019). Determinants of vietnamese listed firm performance: Competition, wage, ceo, firm size, age, and international trade. *Journal of Risk and Financial Management*, 12(2):62.

Wang, G., DeGhetto, K., Ellen, B. P., and Lamont, B. T. (2019). Board antecedents of ceo duality and

the moderating role of country-level managerial discretion: a meta-analytic investigation. *Journal of Management Studies*, 56(1):172–202.

Wang, Y. et al. (2020). Corporate governance mechanisms and firm performance: evidence from the emerging market following the revised cg code. *Corporate Governance: The international journal of business in society*, 20(1):158–174.

Wellalage, N. H. and Locke, S. (2013). Women on board, firm financial performance and agency costs. *Asian Journal of Business Ethics*, 2:113–127.

Wiengarten, F., Lo, C. K., and Lam, J. Y. (2017). How does sustainability leadership affect firm performance? the choices associated with appointing a chief officer of corporate social responsibility. *Journal of business ethics*, 140:477–493.

Wiley, C. and Monllor-Tormos, M. (2018). Board gender diversity in the stemf sectors: the critical mass required to drive firm performance. *Journal of Leadership Organizational Studies*, 25(3):290–308.

Xie, J., Nozawa, W., and Managi, S. (2020). The role of women on boards in corporate environmental strategy and financial performance: A global outlook. *Corporate Social Responsibility and Environmental Management*, 27(5):2044–2059.

Yasser, Q. R. (2012). Affects of female directors on firms performance in pakistan. *Modern Economy*, pages 817–825.

Zemzem, A. and Kacem, O. (2014). Risk management, board characteristics and performance in the tunisian lending institutions. *International Journal of Finance & Banking Studies*, 3(1):186–200.

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

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

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

Appendices

A Supplementary Figures

Figure A1: Time trend of publishing articles included in the dataset

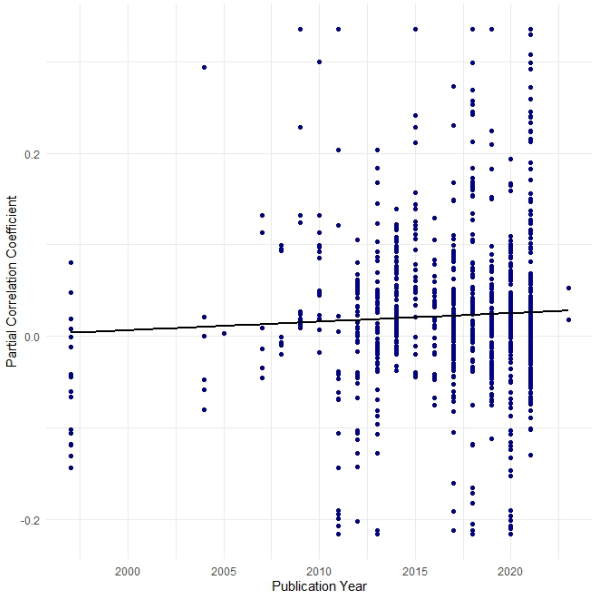


Figure A2: Correlation Matrix

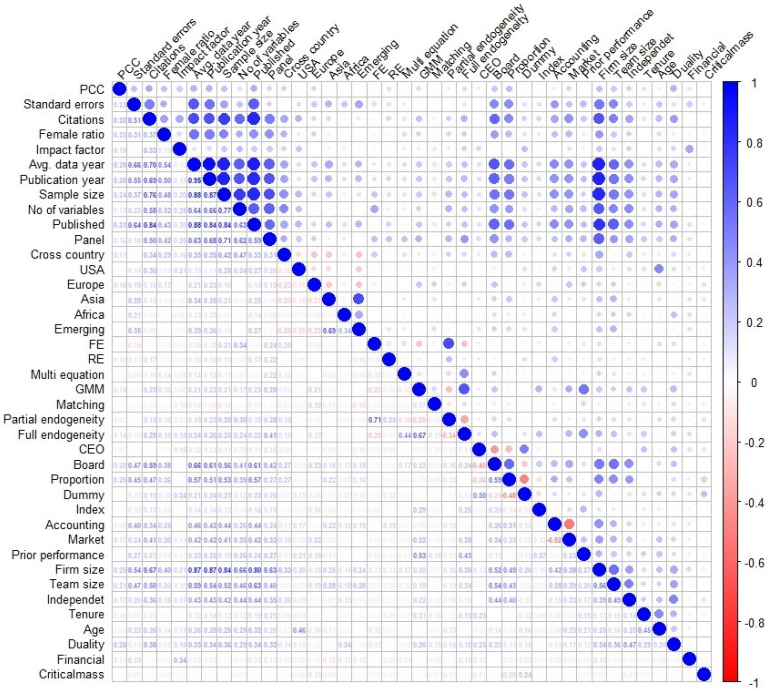


Figure A3: Variation of calculated PCCs across individual studies

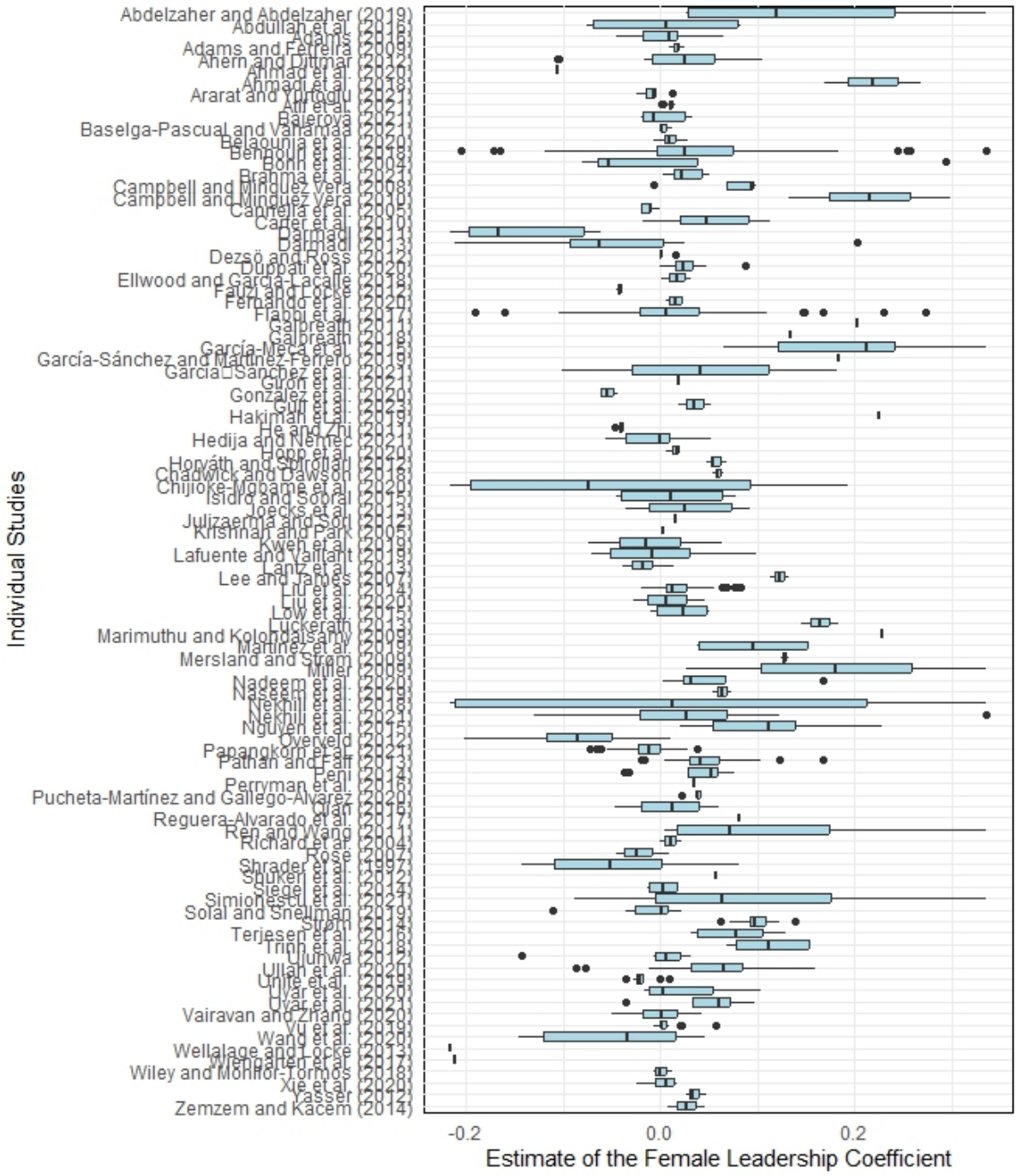


Table A1: Studies identified for analysis

Author (year)	
Abdelzaher and Abdelzaher (2019)	Lantz et al. (2012)
Abdullah and Ismail (2016)	Lee and James (2007)
Adams (2016)	Liu et al. (2020)
Adams and Ferreira (2009)	Liu et al. (2014)
Ahern and Dittmar (2012)	Low and Whiting (2015)
Ahmad et al. (2020)	Lückerath-Rovers (2013)
Ahmadi et al. (2018)	Marimuthu and Kolandaisamy (2009)
Ararat and Yurtoglu (2021)	Martínez and Rambaud (2019)
Atif et al. (2021)	Mersland and Strøm (2009)
Bajerova (2021)	Miller and Triana (2009)
Baselga-Pascual and Vahamaa (2021)	Nadeem et al. (2020)
Belaounia et al. (2020)	Naseem et al. (2019)
Bennouri et al. (2018)	Nekhili et al. (2021)
Bonn et al. (2004)	Nekhili et al. (2018)
Brahma et al. (2021)	Nguyen et al. (2015)
Campbell and Minguez Vera (2010)	Overveld (2012)
Campbell and Minguez-Vera (2008)	Papangkorn et al. (2021)
Cannella Jr et al. (2008)	Pathan and Faff (2013)
Carter et al. (2010)	Peni (2014)
Darmadi (2011)	Perryman et al. (2016)
Darmadi (2013)	Pucheta-Martínez and Gallego-Álvarez (2020)
Dezsö and Ross (2012)	Qian (2016)
Duppati et al. (2020)	Reguera-Alvarado et al. (2017)
Ellwood and Garcia-Lacalle (2018)	Ren and Wang (2011)
Fauzi and Locke (2012)	Richard et al. (2004)
Fernando et al. (2020)	Rose (2007)
Flabbi et al. (2017)	Shrader et al. (1997)
Galbreath (2011)	Shukeri et al. (2012)
Galbreath (2018)	Siegel et al. (2014)
García-Meca et al. (2015)	Simionescu et al. (2021)
García-Sánchez and Martínez-Ferrero (2019)	Solal and Snellman (2019)
García-Sánchez et al. (2021)	Strøm et al. (2014)
Girón et al. (2021)	Terjesen et al. (2016)
González et al. (2020)	Trinh et al. (2018)
Gull et al. (2023)	Ujunwa (2012)
Hakimah et al. (2019)	Ullah et al. (2020)
He and Huang (2011)	Unite et al. (2019)
Hedija and Němec (2021)	Uyar et al. (2021)
Hopp et al. (2020)	Uyar et al. (2020)
Horváth et al. (2012)	Vairavan and Zhang (2020)
Chadwick and Dawson (2018)	Vu et al. (2019)
Chijoke-Mgbame et al. (2020)	Wang et al. (2020)
Isidro and Sobral (2015)	Wellalage and Locke (2013)
Joecks et al. (2013)	Wiengarten et al. (2017)
Julizaerma and Sori (2012)	Wiley and Monllor-Tormos (2018)
Krishnan and Park (2005)	Xie et al. (2020)
Kweh et al. (2019)	Yasser (2012)
Lafuente and Vaillant (2019)	Zemzem and Kacem (2014)

IES Working Paper Series

2024

1. Nino Buliskeria, Jaromir Baxa, Tomáš Šestořád: *Uncertain Trends in Economic Policy Uncertainty*
2. Martina Lušková: *The Effect of Face Masks on Covid Transmission: A Meta-Analysis*
3. Jaromir Baxa, Tomáš Šestořád: *How Different are the Alternative Economic Policy Uncertainty Indices? The Case of European Countries.*
4. Sophie Ghvanidze, Soo K. Kang, Milan Ščasný, Jon Henrich Hanf: *Profiling Cannabis Consumption Motivation and Situations as Casual Leisure*
5. Lorena Skufi, Meri Papavangjeli, Adam Gersl: *Migration, Remittances, and Wage-Inflation Spillovers: The Case of Albania*
6. Katarina Gomoryova: *Female Leadership and Financial Performance: A Meta-Analysis*

All papers can be downloaded at: http://ies.fsv.cuni.cz

Univerzita Karlova v Praze, Fakulta sociálních věd
Institut ekonomických studií [UK FSV – IES] Praha 1, Opletalova 26
E-mail : ies@fsv.cuni.cz http://ies.fsv.cuni.cz