Do Female Directors Enhance ESG Performance? A Meta-Analysis*

Karolina Hozova[a], Tomas Havranek[a,b,c], and Zuzana Irsova[a,c]
[a]Institute of Economic Studies, Faculty of Social Sciences, Charles University, Prague
[b]Centre for Economic Policy Research, London
[c]Meta-Research Innovation Center, Stanford

July 2, 2026

Abstract

Appointing more women to corporate boards is widely expected to also raise firms' environmental, social, and governance (ESG) performance, letting one decision serve two goals. We provide the first meta-analysis of this relationship, drawing on 533 estimates from 106 studies that measure ESG performance with Bloomberg or LSEG ratings. The average reported effect of a one-percentage-point increase in board gender diversity is about 0.28 ESG points, but much of it does not survive scrutiny. Correcting for publication bias with a battery of linear and non-linear methods lowers the effect to between roughly 0.08 and 0.17 points; a best-practice estimate that also imposes sound study design puts it near 0.12 for most of the world, markedly higher for the Middle East, and essentially zero, if anything slightly negative, for the Southeast Asian markets that dominate the Asian evidence. The differences that remain across studies are systematic, driven mainly by geography and by the choice of estimation method rather than by the ESG-rating provider or the controls a study includes. Board gender diversity may be well worth pursuing on its own merits, but the evidence that it reliably raises ESG scores is weaker than the published record suggests.

*Corresponding author: Karolina Hozova, karolina.hozova@fsv.cuni.cz. Hozova acknowledges support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 870245

1 Introduction

Companies face pressure on two fronts at once. Investors, regulators, and the public want firms to put more women in the boardroom, and they want firms to do better on environmental, social, and governance (ESG) measures. For a board deciding where to spend its effort, a tempting shortcut suggests itself: if women directors genuinely improve a firm's ESG performance, appointing more of them advances both goals together. One appointment, two problems solved.

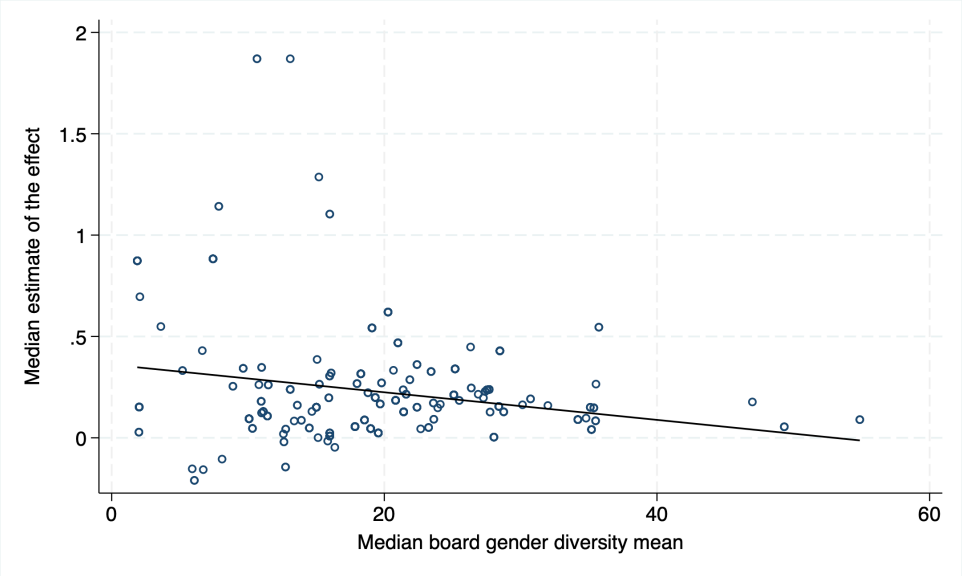
The idea is not far-fetched. A large body of work in psychology and management argues that women bring distinct priorities to the boardroom. Theories of gender socialization hold that women are, on average, more attuned to the welfare of others and less willing to tolerate unethical behavior (Gilligan, 1977; Boulouta, 2013; Williams, 2003). Carried into corporate leadership, these tendencies are thought to make female directors more responsive to the concerns of employees, communities, and the environment (Bear et al., 2010; Harjoto et al., 2015; Atif et al., 2021).

The empirical record is far less tidy. Manita et al. (2018), among the most cited studies on the topic, examined US firms between 2010 and 2015 and found an effect indistinguishable from zero. Three years later, Shakil et al. (2021) revisited the question on US banks with more recent data and reported a clear positive effect of roughly half a point; the two papers were soon cited side by side, one appearing to overturn the other's null. Similar disagreements run across regions and sectors, from Latin America (Husted and de Sousa-Filho, 2019) and the Middle East (Issa et al., 2022) to energy (Shahbaz et al., 2020), healthcare (Uyar et al., 2021), and mining (Wang et al., 2022). Across the 533 estimates we collect, the reported effect of a one-percentage-point increase in board gender diversity ranges from -2.8 to 5.9 ESG points in the raw data.

Part of this disagreement tracks the level of female board representation itself. Figure 1 plots each study's median effect against the average share of women on its sample's boards. The downward slope is stark: studies from contexts where women hold few board seats (often firms in the Middle East or other emerging markets, where the sample mean rarely exceeds ten percent) report the largest effects, sometimes above one ESG point per percentage point, while studies from high-representation settings such as Northern Europe or the United States cluster near zero. A literature that looks at first glance as though it measures one thing turns out to be measuring very different margins in very different institutional settings. How readily women reach board seats depends on regulation, investor pressure, and cultural norms as much as on the sector, so a single pooled estimate can mask sharply different local relationships (Carrasco et al., 2015).

What explains this spread, and what is the effect beneath it? A literature with an appealing story, mixed findings, and wide variation is fertile ground for publication bias: the tendency for results of the expected sign and conventional significance to be reported,

Figure 1: Heterogeneity in the literature



Notes: Each circle represents one primary study. The vertical axis shows the study's median estimate of the effect of a one-percentage-point increase in board gender diversity on ESG scores. The horizontal axis shows the median sample mean of board gender diversity in that study. The fitted line is from an OLS regression.

cited, and published more readily than null results (Stanley, 2005), a pattern documented across economics from minimum-wage research (Card and Krueger, 1995; Doucouliagos and Stanley, 2009) and fiscal policy (Heinemann et al., 2018) to the elasticity of factor substitution (Gechert et al., 2022), the growth effects of remittances (Cazachevici et al., 2020), and behavioral economics (Havránek et al., 2017). This literature looks especially exposed. A null result on so appealing a hypothesis makes for a far less compelling paper than a positive one, so positive and significant estimates likely reach print more easily, and the average effect circulating in the literature may overstate whatever genuine relationship exists.

We provide the first assessment of how much of this evidence survives such scrutiny. We hand-collect 533 estimates from 106 studies that measure ESG performance with Bloomberg or LSEG ratings, restricting the sample to this common scale so that effect sizes are comparable across studies. These ratings are third-party assessments and need not track a firm's underlying conduct, so our results speak to rated ESG performance rather than to behavior we observe directly. To separate any genuine effect from selective reporting, we apply a battery of linear and non-linear corrections: the FAT-PET regression of Stanley (2005, 2008), the weighted average of adequately powered estimates of Stanley et al. (2017), the selection model of Andrews and Kasy (2019), the stem-based method of Furukawa (2019), the endogenous kink model of Bom and Rachinger (2019), and the p-uniform* estimator of van Aert and van Assen (2026). We then ask what drives the disagreement, using Bayesian model averaging over 24 variables (the standard error

and 23 study characteristics) to avoid betting the analysis on a single specification.

The raw literature implies that a one-percentage-point increase in board gender diversity raises ESG scores by about 0.28 points. Much of that reflects selective reporting: correcting for publication bias alone brings the typical effect down to between 0.08 and 0.17 points, depending on the method. A best-practice estimate that additionally imposes sound study design (a panel setup and no publication bias) puts the effect near 0.12 for most of the world, markedly higher for Middle Eastern firms, and essentially zero, if anything slightly negative, for the Southeast Asian markets that dominate our Asian evidence. The heterogeneity across studies is systematic rather than random, driven mainly by geography and by a few study characteristics, especially panel-data use and estimation method; the headline magnitude is largely a product of how the literature was built rather than a reliable measure of how firms behave.

Related meta-analyses examine women on boards and firm financial performance (Post and Byron, 2015), women directors and corporate social performance (Byron and Post, 2016), and gender diversity and corporate disclosure (Wu et al., 2022; AlJanadi, 2025). None of them studies third-party ESG ratings on a common scale; Wu et al. (2022), for instance, synthesize 44 papers and find that firms with more women on the board disclose more, not that they earn higher ESG scores. To the best of our knowledge, this is the first meta-analysis of the association between board gender diversity and third-party ESG ratings measured on a comparable Bloomberg or LSEG scale, and the first to test this ESG-ratings literature for publication bias and to identify, rather than assume, what drives the disagreement among its estimates.

The remainder of the paper proceeds as follows. Section 2 describes the dataset of board gender diversity effects. Section 3 explores publication bias. Section 4 examines the drivers of heterogeneity. Section 5 concludes. Appendix A details how studies were selected for inclusion, and Appendix B and Appendix C report robustness checks.

The data and code behind every table and figure are available in an online appendix at meta-analysis.cz/esg. The package includes the full dataset of 533 estimates with a codebook, the Stata and R scripts, and documentation of the order in which they run, so that our coding decisions and each step of the analysis can be inspected and reproduced. We follow the reporting guidelines for meta-analyses in economics of Stanley et al. (2013), Havránek et al. (2020), and Iršová et al. (2024).

2 Data

Our analysis starts with an original data compilation. Altogether, we collect 533 estimates of the effect of board gender diversity on ESG scores from 106 research papers. These primary studies are identified using Google Scholar, which searches the full text of articles rather than just titles, abstracts, and keywords. To keep the search process replicable,

we rely on a single search query. In addition, we complement the identification of studies by screening references of the primary studies identified through Google Scholar search. Backward snowballing works through the reference lists of the studies already included. We do not perform forward snowballing (tracking citations to the identified studies); instead, shortly before finalizing the dataset we ran a second Google Scholar search to capture studies published since the first search. Complete details of the identification strategy and the specific search query are presented in Figure A.1 in Appendix A.

For comparability, we consider only studies that examine the impact of board gender diversity, measured as the ratio of female directors on the board, on ESG scores rated by Bloomberg or London Stock Exchange Group (LSEG), both reported on the same 0 to 100 scale (the provider floor is 0.01 rather than exactly zero). ESG scores can differ across providers (Dorfleitner et al., 2015; Berg et al., 2022); we show below that our results do not depend on whether a study uses Bloomberg or LSEG ratings. To keep the effect size comparable across studies, we exclude estimates that omit one of the three ESG pillars, log-transform the ESG score, or discount it by a controversy score (Shakil et al., 2021), keeping a study's comparable ESG-score estimates when it also reports them. Similarly, studies that use statistical measures such as the Blau or Shannon index (Abdullah et al., 2024; Gangi et al., 2021) and studies that focus solely on female CEO or chair leadership (Aabo and Giorici, 2023; Harjoto and Wang, 2020; Li et al., 2023; Liu et al., 2024), proportion of women in top management teams (Fu et al., 2023), dummies for female board representation (Chebbi and Ammer, 2022) or report only the number of women on board (Kravchenko et al., 2023) are not included. On the other hand, we do not exclude studies based on their publication form (Stanley, 2001), provided they report a measure of uncertainty such as standard errors, confidence intervals, or p-values. Our final sample therefore includes not only standard peer-reviewed journal articles but also working papers and master's theses. Table 1 presents the final list of primary studies used in our meta-analysis.

Alongside the individual effect estimates and their uncertainty measures from the primary studies in our list, we further hand-collected other variables to examine the systematic heterogeneity among the reported coefficients. These variables encompass estimation and publication characteristics, specific data features and information on spatial variation.

During data collection, we made a few adjustments to ensure that our dataset contains comparable effect estimates. First, some studies explored a non-linear relationship between board gender diversity and ESG scores by including a quadratic term (Birindelli et al., 2018). To handle the presence of two related estimates, we follow the methodology of Žigraiova and Havránek (2016) and linearize the effect. Second, four studies report standardized effects (Dakhli, 2021; Dang et al., 2023a; Kamran et al., 2023; Khemakhem et al., 2023). Standardized estimates from these studies are recomputed to raw effects us-

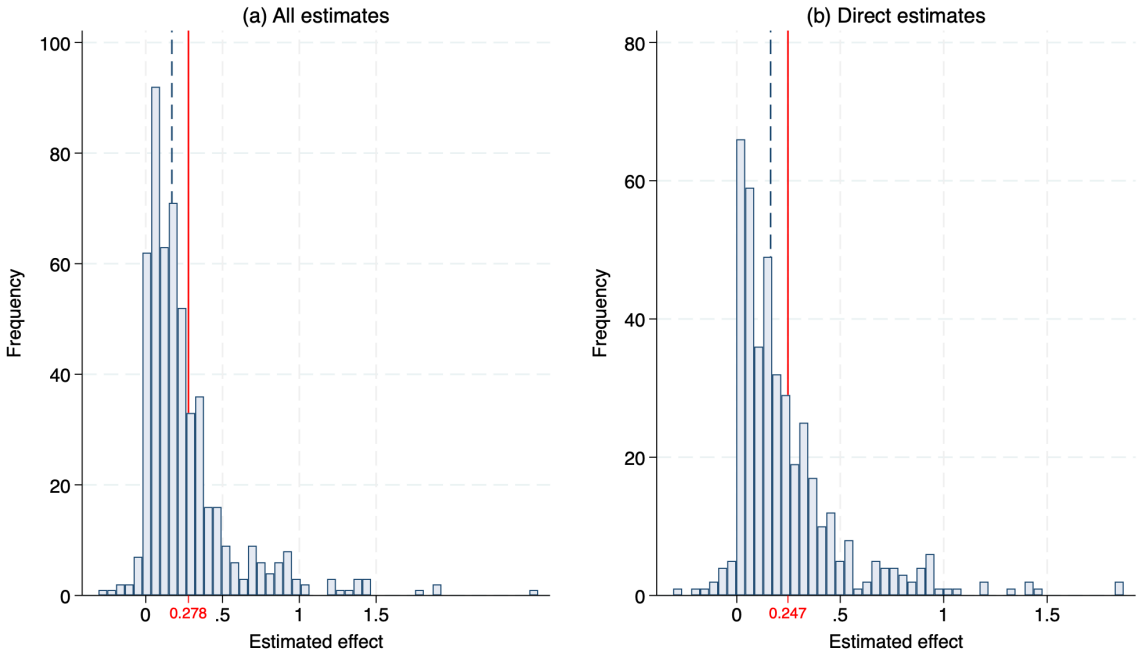
Table 1: Primary studies used

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

Notes: The table lists all primary studies identified through the search strategy described in Section 2 (Google Scholar and snowballing). The last study was added on December 16, 2024.

ing the standard deviation ratio. Third, some studies employ interaction terms between board gender diversity and other variables, such as common law tradition (Alkhawaja et al., 2023), the critical mass of women on the board (Birindelli et al., 2018) or ESG controversies (Shakil et al., 2021). In line with the approach by Cazachevici et al. (2020), we compute the average marginal effects by applying the delta method to derive the corresponding standard errors. Whenever any of the mentioned transformations is applied, we document it in our dataset, so we can exclude such estimates from our robustness checks. For some studies, the lack of summary statistics or uncertainty measures prevents us from applying the delta method or effect standardization (Dang et al., 2023b; Nuhu and Alam, 2024). Also, in cases where the mean value for the variable included in

Figure 2: Distribution of gender diversity effects



Notes: The figure presents a distribution of the estimated effects of board gender diversity on ESG scores as reported across individual studies. Panel (a) shows all estimates; panel (b) restricts to direct estimates. In each panel the solid vertical line marks the mean reported effect and the dashed vertical line the median; across all estimates the mean corresponds to a 0.278-point increase in ESG rating following a 1-percentage-point increase in female board representation.

the interaction term is too high and produces extremely large average marginal effects, we do not include them in our dataset (Giannarakis, 2013).

A few studies reported p-values or standard errors of zero, which required imputation. For reproducibility reasons, zero standard errors were replaced with 0.0004 (Bruna et al., 2021; Gerged et al., 2023) and zero p-values (Bhatia and Marwaha, 2022; Martínez et al., 2022; Nadeem et al., 2017; Setiani and Novitasari, 2024) or significance levels of 1% (Miranda et al., 2023) were replaced by 0.0001. For studies reporting only that an estimate is significant at the 5% level (Fahad and Rahman, 2020; Kamaludin et al., 2022), we set the t-statistic to 2.27, the midpoint between the two-tailed 5% and 1% critical values. Estimates with imputed standard errors or p-values are also marked in the dataset and excluded from the robustness check.

Finally, despite rigorous data cleaning, some outliers in the estimated effect sizes, along with their associated standard errors, remain. To keep these observations informative without letting them skew the results, we winsorize the estimates and their standard errors at the 1% level, replacing the five most extreme observations in each tail of each variable, so that the largest retained effect is about 1.9 ESG points. Appendix B reports all publication-bias tests on the raw, unwinsorized data. The nonlinear and precision-

weighted corrections remain small, positive, and statistically significant in this check, but on the raw data the linear FAT-PET slope loses significance in the unweighted and study-weighted specifications, whose corrected means rise to about 0.25–0.31; we therefore rest the corrected-effect claim on the estimators that survive both.

The final dataset includes 533 estimates derived from 106 primary studies. The studies

Table 2: Board gender diversity effects in different contexts

		Weighted			Unweighted		
	No. of estimates	Mean	95% conf. int.		Mean	95% conf. int.	
All estimates	533	0.278	0.245	0.311	0.269	0.241	0.298
Direct estimates	426	0.226	0.197	0.254	0.246	0.218	0.274
Non-direct estimates	107	0.501	0.388	0.613	0.364	0.275	0.452
Data type							
Data: panel	512	0.258	0.227	0.289	0.263	0.235	0.290
Data: cross-sectional	21	0.544	0.262	0.825	0.434	0.144	0.723
ESG data: LSEG	329	0.297	0.258	0.336	0.270	0.236	0.303
ESG data: Bloomberg	204	0.245	0.186	0.304	0.269	0.216	0.322
Publication status							
Published	506	0.290	0.254	0.325	0.271	0.241	0.301
Unpublished	27	0.171	0.098	0.245	0.239	0.151	0.327
Design of the analysis							
Endogeneity control: poor	435	0.283	0.246	0.320	0.275	0.244	0.306
Endogeneity control: proper	98	0.259	0.182	0.335	0.245	0.172	0.318
Control variables							
Firm size control: yes	474	0.299	0.263	0.335	0.284	0.253	0.316
Firm size control: no	59	0.123	0.067	0.179	0.149	0.099	0.199
Board independence control: yes	357	0.277	0.238	0.316	0.267	0.233	0.302
Board independence control: no	176	0.280	0.217	0.342	0.273	0.221	0.326
CSR committee control: yes	206	0.234	0.195	0.273	0.259	0.216	0.301
CSR committee control: no	327	0.302	0.255	0.349	0.276	0.238	0.315
Spatial variation							
Region: global	177	0.290	0.226	0.354	0.219	0.167	0.270
Region: Europe	182	0.202	0.181	0.224	0.232	0.208	0.257
Region: USA	42	0.316	0.198	0.433	0.295	0.197	0.394
Region: Asia	43	0.174	0.056	0.293	0.177	0.071	0.283
Region: Middle East	30	0.577	0.394	0.760	0.657	0.459	0.855
Region: other	59	0.462	0.295	0.629	0.388	0.277	0.499
Market: developed	274	0.274	0.234	0.314	0.257	0.229	0.286
Market: emerging	98	0.278	0.194	0.362	0.411	0.318	0.504
Market: mixed	161	0.290	0.216	0.364	0.204	0.147	0.261
Sector: financial	37	0.294	0.206	0.381	0.385	0.264	0.506
Sector: non-financial	152	0.197	0.173	0.221	0.216	0.186	0.246
Sector: mixed	344	0.313	0.264	0.363	0.281	0.240	0.321
Estimation techniques							
Method: linear	186	0.263	0.215	0.311	0.264	0.220	0.307
Method: panel (FE or RE)	215	0.226	0.185	0.267	0.238	0.196	0.280
Method: IV	33	0.656	0.437	0.875	0.423	0.253	0.592
Method: GMM	42	0.334	0.167	0.502	0.282	0.141	0.424
Method: other	57	0.224	0.139	0.308	0.309	0.219	0.399

Notes: The table displays subgroup summary statistics of the estimated effects of board gender diversity on ESG scores. Weighted means give each study equal weight; unweighted means give each estimate equal weight.

included in the sample were published between 2013 and 2024. Only 16 of the 106 studies were published before 2021, reflecting the topic's recent rise. Figure 2 shows the distribution of estimated effects. The left-hand panel shows all estimates, while the right-hand panel restricts to direct estimates (those reported without transformation or imputation). Both histograms show a heavy-tailed distribution with a peak around zero and positive skewness.

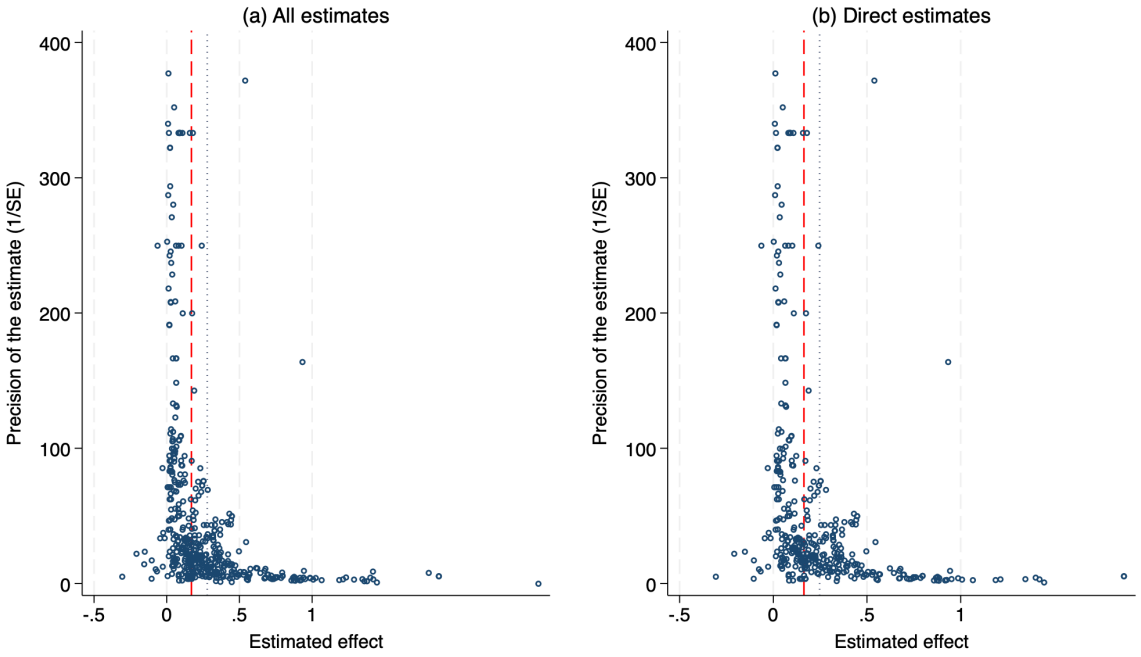
The box plot in Figure A.2 in Appendix A illustrates the heterogeneity of the estimates in different regions. Consistent with the histograms presented in Figure 2, most of the estimates are positive. Some regions, notably Latin America and the Middle East, exhibit greater variation, with wider interquartile ranges and longer whiskers.

Table 2 puts numbers on these patterns, reporting both weighted and unweighted means across categories. A weighted mean gives each study equal total weight (the inverse of its number of estimates), whereas an unweighted mean gives each estimate equal weight. Weighted means are preferred because our dataset includes primary studies that report disproportionately many estimates (Alkhawaja et al., 2023), whereas other primary studies report only a single estimate (Waterstraat et al., 2021; Yarram and Adapa, 2021). Although the overall sample means do not differ dramatically, some subsample means vary much more. The overall estimated effect of female board participation on firms' ESG is around 0.278 (weighted) and 0.269 (unweighted). Both means suggest a slight but positive relationship between board gender diversity and ESG scores.

A closer look at the mean values in different contexts shows that studies using panel data and Bloomberg's ESG data yield more conservative estimates. In contrast, the raw subgroup means are somewhat higher among studies that omit a corporate social responsibility committee control (0.30 versus 0.23), while adequacy of endogeneity control makes little difference (0.28 versus 0.26); Section 4 shows that, of the patterns just described, only the panel-data one survives once other study characteristics are accounted for. The effect also seems more pronounced in the Middle East than in European or US countries. Publication status also matters: the effect is more pronounced in published studies than in unpublished ones.

At this stage, whether the literature is subject to publication bias remains an open question. Although the weighted mean indicates substantial differences in the estimated effect between published and unpublished studies, simple averages can obscure the underlying distortions in the reported findings (Ioannidis et al., 2017). The following section therefore investigates in more detail whether the empirical evidence on board gender diversity and ESG scores is subject to publication bias.

Figure 3: Funnel plot of reported estimates



Notes: Panel (a) shows all estimates; panel (b) restricts to direct estimates. In the absence of publication bias, the most precise estimates are expected to cluster around the mean effect. Less precise estimates should be symmetrically distributed around it. The figure suggests an asymmetry, with the most precise estimates located between zero and the mean effect (solid vertical line). The dashed line represents the median effect. Extreme outliers are omitted from the figure but remain included in all statistical tests.

3 Publication Bias

Publication bias shows up in a meta-analysis as a correlation between reported estimates and their standard errors: when significant results of the expected sign are more likely to be published, less precise studies must report larger effects to clear the bar for significance (Stanley, 2005). We look for this pattern first visually, with a funnel plot, and then test for it formally.

To assess potential publication selection bias, we employ a widely used visual diagnostic tool for meta-analyses, the funnel plot (Egger et al., 1997). This plot shows individual effect estimates on the horizontal axis against their precision, the inverse standard error, on the vertical axis.

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

Applying a funnel plot to our sample of estimates produces the slightly skewed inverted funnel depicted in Figure 3. The most precise estimates cluster to the left of the mean effect, signaling the presence of publication bias. The funnel plot for direct, non-imputed estimates confirms the same two features: negative estimates are slightly underrepresented, and the most precise estimates lie between zero and the sample mean of the estimated effect. However, the funnel plot asymmetry must be interpreted with caution, as it can also be driven by other factors such as heterogeneity between studies, differences in methodology, or data quality issues (Egger et al., 1997). Although this graphical method suggests a right-skewed funnel plot, we complement it in the following section with formal statistical tests for asymmetry.

Egger et al. (1997) propose a linear approach to test for publication bias, which measures the symmetry of the funnel plot. This method, also called the Egger regression, examines the correlation between reported estimates and their standard errors. If publication bias is present, the correlation will be significantly different from zero (Stanley, 2005).

Following the now-standard calibration of Doucouliagos and Stanley (2013) for gauging the severity of publication selection, if the funnel asymmetry test (FAT) is not statistically significant or the absolute value of β_1 is less than 1, the degree of publication bias is classified as little to modest. Publication bias is considered substantial if FAT is significant and the absolute value of β_1 ranges between 1 and 2. Consequently, if β_1 exceeds 2 and FAT is significant, publication bias is classified as severe.

Egger regression can be estimated using various model specifications. We begin with ordinary least squares. Second, we estimate a model that uses the between-study variance. A model that employs within-study variance is not estimated, as some primary studies provide only one effect estimate. Third, we employ weighting schemes commonly applied in meta-analyses, following Gechert et al. (2022) and Havránek et al. (2018): weighting by the inverse of the number of estimates per study and by precision of estimates. The first weighting gives each study equal weight regardless of how many estimates it reports, while the latter one ensures that less weight is given to less precise estimates and the heteroskedasticity from FAT is eliminated.

We address the potential heteroskedasticity in the FAT-PET framework by clustering standard errors at the study level. The standard assumption of independently and identically distributed error terms is likely violated because of within-study correlations among reported estimates. Clustering standard errors at the study level allows us to account for this intra-study dependence while maintaining the assumption of independence across different studies. Although our sample meets the minimum threshold for valid inference, the presence of unequal cluster sizes may still introduce bias, as noted by MacKinnon and Webb (2017). To mitigate this issue, we adopt the wild cluster bootstrap of Roodman et al. (2019), which is particularly robust to cluster imbalance. In our data the imbalance

is moderate: the number of estimates per study ranges from one to 28 (median two), and no single study contributes more than 5.3% of the sample, which limits the influence any one cluster can exert on the bootstrap inference. We report 95% confidence intervals based on this procedure for all specifications, with the exception of the between-effects estimation at the study level.

Finally, because estimating the Egger regression might suffer from endogeneity between the estimates and their standard errors, we follow Iršová et al. (2025) and instrument the reported variance using the meta-analysis instrumental variable estimator (MAIVE). Bias usually arises from random sampling errors and joint computation of estimates and their standard errors, both potentially influenced by the estimation method chosen in the primary study. An instrument satisfying both the relevance and exogeneity assumptions is the inverse of the number of observations. Studies that use larger samples should produce smaller standard errors, and sample sizes should not be correlated with the chosen estimation method.

Panel A of Block 1 in Table 3 summarizes the results from the model specifications described above. All four regression-based specifications show positive, substantial to severe publication bias significant at the 1% level; MAIVE, discussed below, does not identify a separate bias coefficient and its corrected mean is statistically indistinguishable from zero given the weak instrument. Compared to the weighted and unweighted means from the primary studies (0.278 and 0.269, see Table 2), the bias-corrected mean is notably lower, ranging from 0.076 to 0.113. MAIVE pulls the mean toward zero beyond the bias correction; however, given the reported F-statistics, the instrument is weak, so its point estimate remains unreliable.[1]

Although funnel asymmetry and precision-effect tests are widely used to detect publication bias and to estimate the true underlying effect (Stanley, 2008; Stanley and Doucouliagos, 2014), a key limitation of these approaches lies in their assumption of a linear relationship between standard errors and estimates. This assumption may not hold in practice, especially for highly precise estimates: their inherently small standard errors can yield significance even without selective reporting (Stanley et al., 2010), so such estimates are less likely to be influenced by publication bias. As a result, linear models may exaggerate the extent of bias and underestimate the true effect. To address this issue, we complement the linear FAT-PET approach with a set of non-linear estimators.

As a first step in our non-linear checks, we implement the weighted average of adequately powered (WAAP) estimator, introduced by Stanley et al. (2017). This approach

[1]The figures in curly brackets are Anderson and Rubin (1949) 95% confidence intervals, which are robust to weak identification and are obtained by inverting the Anderson–Rubin test rather than from the point estimate's standard error. They therefore need not be centered on, or even contain, the MAIVE point estimate; in the full-sample and direct-estimate blocks the interval lies entirely above it. We report these intervals for completeness but, given the weak first stage, do not base any conclusion on MAIVE's point estimate.

aims to mitigate the impact of publication bias by retaining only those estimates whose statistical power to detect the (unrestricted) weighted-average effect exceeds 80%. Hence, less precise, underpowered estimates that could potentially inflate bias are excluded. The adequately powered estimates are then aggregated using optimal inverse-variance weights ($1/SE^2$) to derive a more reliable average effect size. However, a known limitation of WAAP, as noted by Stanley et al. (2017), is its dependency on the presence of sufficiently powered studies within the dataset. When such studies are scarce or absent, WAAP becomes uninformative. A further caveat is that the power screen is benchmarked against the unrestricted weighted-average effect, which is itself not immune to the selection we are trying to correct; we therefore treat WAAP as one input among several rather than as a stand-alone correction. Thus, we complement it with a few other checks.

The selection model of Andrews and Kasy (2019) builds on the assumption that the likelihood of publishing an effect estimate is influenced by its statistical significance. Specifically, the model posits that this probability shifts once the estimate surpasses certain t-statistic thresholds. Using maximum likelihood estimation, the model calculates the publication probabilities for different data windows defined by these critical t-statistic thresholds. Then it identifies which estimates are underrepresented within these windows and assigns them greater weight, thus adjusting for potential publication bias. This approach resembles the one proposed by Hedges (1992) but offers a more nuanced treatment of the relationship between statistical significance and the likelihood of publication.

We also apply two non-linear methods, the STEM-based method and the Endogenous Kink method, both of which extend the logic behind the "Top 10" approach introduced by Stanley et al. (2010). These approaches assume that a subsample of the most precise

Table 3: Tests for publication bias: full sample and direct estimates

Block 1: Full sample					
Panel A: Linear	OLS	Between Effects	Study Weight	Precision Weight	MAIVE
Publication Bias	1.791***	2.127***	1.837***	2.174***	
(*Standard Error*)	(0.282)	(0.210)	(0.278)	(0.308)	
	[1.175, 2.421]		[1.062, 2.505]	[1.474, 2.859]	
Mean Beyond Bias	0.113***	0.076**	0.110***	0.079***	-0.017
(*Constant*)	(0.020)	(0.031)	(0.026)	(0.020)	(0.098)
	[0.072, 0.156]		[0.057, 0.161]	[0.031, 0.125]	{0.039, 0.191}
First-stage robust F-stat					6.76
Panel B: Nonlinear	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		$P = 0.270$		1.806**	
		(0.042)		(0.833)	
Effect Beyond Bias	0.087***	0.113***	0.173***	0.084***	0.174***
	(0.006)	(0.010)	(0.013)	(0.004)	(0.029)
# of estimates	220		106		
% of information			100%		
Observations	533	533	533	533	533
Studies	106	106	106	106	106

Table 3: Tests for publication bias: full sample and direct estimates (continued)

Block 2: Direct estimates					
Panel A: Linear	OLS	Between Effects	Study Weight	Precision Weight	MAIVE
Publication Bias	1.957***	1.721***	2.098***	2.312***	
(*Standard Error*)	(0.323)	(0.312)	(0.294)	(0.354)	
	[1.077, 2.787]		[1.236, 3.078]	[1.454, 3.128]	
Mean Beyond Bias	0.104***	0.101***	0.098***	0.078***	0.062
(*Constant*)	(0.019)	(0.033)	(0.020)	(0.019)	(0.046)
	[0.064, 0.141]		[0.059, 0.139]	[0.029, 0.127]	{0.103, 0.190}
First-stage robust F-stat					10.32
Panel B: Nonlinear	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		$P = 0.284$		2.043**	
		(0.053)		(1.000)	
Effect Beyond Bias	0.085***	0.115***	0.170***	0.081***	0.153***
	(0.007)	(0.013)	(0.013)	(0.005)	(0.036)
# of estimates	187		94		
% of information			99.9%		
Observations	426	426	426	426	426
Studies	95	95	95	95	95

Notes: Block 1 is the full sample; Block 2 restricts to direct estimates. *Panel A:* FAT-PET regression $E_{is} = \beta_0 + \beta_1 \cdot \text{SE}(E_{is}) + \varepsilon_{is}$; where β_1 is publication bias and the constant β_0 is the mean beyond bias. Cluster-robust standard errors are in parentheses, wild-bootstrap CIs in square brackets (Roodman et al., 2019) and the Anderson and Rubin (1949) 95% CI in curly brackets for MAIVE by Iršová et al. (2025) that corrects for spurious precision. *Panel B:* WAAP denotes weighted average of adequately powered estimates (Stanley et al., 2017); Selection model denotes the technique due to Andrews and Kasy (2019); Stem denotes the stem-based technique (Furukawa, 2019); Kink denotes the endogenous kink model (Bom and Rachinger, 2019); p-uniform* denotes the technique due to van Aert and van Assen (2026). Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

estimates is less prone to publication bias, thus providing a more accurate estimate of the true underlying effect. The STEM-based method, developed by Furukawa (2019), focuses on the "stem" of the funnel plot, selecting the most precise estimates by minimizing the mean squared error. The Endogenous Kink method due to Bom and Rachinger (2019) fits a piecewise linear meta-regression of estimates against their standard errors. The regression consists of the flat segment that represents the most precise estimates unaffected by standard errors and a positively sloped segment that indicates the correlation between standard errors and estimates impacted by publication bias. The intersection of these two segments, known as the "kink", serves as the threshold for identifying the most reliable estimates. Both methods endogenously determine the proportion of precise estimates to include in the final calculation of the average effect. These should be large enough to ensure efficiency, but small enough to exclude imprecise estimates.

As a final non-linear check that avoids regressing estimates on their standard errors, we employ the p-uniform* method developed by van Aert and van Assen (2026). In our case, the p-uniform* is estimated using the method of moments, which rests on the principle that p-values should follow a uniform distribution. However, publication bias distorts

this distribution. Specifically, statistically significant estimates are over-represented in the published record. Conditional on significance, the p-values implied by the reported estimates should be uniformly distributed only at the true effect size. The objective of the p-uniform* method is therefore to find the effect value at which the conditional distribution of these p-values is uniform; this value is the bias-corrected mean.

The results of the non-linear tests described above are presented in Panel B of Block 1 in Table 3. On average, the estimated effects beyond bias generated by non-linear tests are slightly higher than the ones generated by linear tests, ranging from 0.084 to 0.174.

The primary robustness check reported in Block 2 of Table 3 excludes 107 estimates on which any form of transformation was performed and estimates that had their uncertainty statistics imputed. For MAIVE the first-stage F rises to 10.32, at the conventional threshold, and the weak-identification-robust Anderson–Rubin interval (0.103 to 0.190) lies entirely above zero, corroborating a small positive effect while the point estimate remains uninformative. We repeat the tests on two additional samples, one excluding Middle East observations and one using raw unwinsorized data, both reported in Table B.1 in Appendix B. Excluding Middle East observations closely replicates the full-sample findings. On raw, unwinsorized data the linear evidence of publication bias weakens and some linear corrected means rise, whereas the non-linear estimates remain small and positive (Table B.1). The F-statistic falls below the conventional threshold of 10 in both cases, indicating a weak instrument. Finally, because our sample pools two ESG-rating providers, we test whether the results depend on the provider. According to Berg et al. (2022) and Dorfleitner et al. (2015), using ESG scores from different data providers may yield different results. An interacted funnel-asymmetry test that allows both the slope and the intercept to differ for Bloomberg-rated estimates finds neither difference statistically significant, so we cannot reject provider invariance (Table B.2 in Appendix B).

4 Heterogeneity

Heterogeneity in the literature may influence the correlation between estimates and standard errors, originally attributed to publication bias. We therefore test whether the publication-bias results survive the inclusion of control variables that reflect study design, and identify which of these variables systematically explain heterogeneity among the reported estimates. According to Adams et al. (2015), these could be the use of different samples, time windows or empirical methods. To explore the sources of heterogeneity among the reported estimates, we codify variables organized into six categories: data characteristics, publication characteristics, estimation techniques, design of the analysis, control variables, and spatial variation.

The resulting variables are described in Table 4. While each could influence the re-

ported board gender diversity effects, only a few are likely to matter consistently. Adding all variables into one model would likely produce very imprecise estimates, even for the key ones. In contrast, selecting a single *best* model among all possible combinations would not only be arbitrary, but it would also ignore the uncertainty that inherently comes with such a decision.

Bayesian model averaging (BMA) addresses this directly. Instead of relying on a single specification, it treats the model space itself as uncertain within a probabilistic framework. It does so by evaluating all possible combinations of explanatory variables, each of them forming its own potential model, and assigning to each one a posterior model probability (PMP) based on how well it explains the data (Raftery et al., 1997). This means that instead of choosing one model and discarding the rest, all models are weighted by how likely they are to be the *true* model. The result is a transparent approach that acknowledges, rather than suppresses, the uncertainty inherent in model selection (Eicher et al., 2011; Havránek et al., 2015).

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

Table 4: Overview and descriptive statistics of contextual variables

Variable	**Description**	**Mean**	**SD**	**WM**
Estimate	= estimated effect	0.269	0.338	0.278
Standard error	= estimated standard error of the estimate	0.087	0.124	0.095
Data characteristics				
Sample size	= logarithm of the sample size (no. of observations)	7.327	1.732	6.896
Data: panel	= 1 if panel data is used for estimation	0.961	0.195	0.931
ESG data: LSEG	= 1 if LSEG's ESG data is employed	0.617	0.487	0.623
Average data year	= logarithm of the average data year	7.609	0.001	7.609
Average board gender diversity	= logarithm of the average sample board gender diversity	2.795	0.641	2.798
Publication characteristics				
Published	= 1 if published in a journal	0.949	0.219	0.896
Unpublished	= 1 if unpublished working paper (reference category for publication status)	0.051	0.220	0.104
Number of citations	= logarithm of total citations	2.871	1.538	2.930
Design of the analysis				
Endogeneity control: poor	= 1 if poor endogeneity control is employed	0.816	0.388	0.784
Endogeneity control: proper	= 1 if proper endogeneity control is employed (reference category for endogeneity control)	0.184	0.388	0.216
Number of variables	= logarithm of the number of variables used in the estimation	2.311	0.466	2.162
Control variables				
Firm size control	= 1 if study controls for size of the firm	0.889	0.314	0.878
Board independence control	= 1 if study controls for independence of the board	0.670	0.471	0.682
CSR committee control	= 1 if study controls for existence of firm's corporate social responsibility committee	0.386	0.487	0.360

Table 4: Overview and descriptive statistics of contextual variables (continued)

Variable	Description	Mean	SD	WM
Spatial variation				
Region: global	= 1 if study employs a global sample of firms	0.332	0.471	0.217
Region: Europe	= 1 if study focuses on European firms	0.341	0.475	0.371
Region: USA	= 1 if study focuses on US firms	0.079	0.270	0.121
Region: Asia	= 1 if study focuses on Asian firms	0.081	0.273	0.137
Region: Middle East	= 1 if study focuses on firms in the Middle East	0.056	0.231	0.057
Region: other	= 1 if study focuses on firms from other countries (reference category for region)	0.111	0.314	0.097
Market: emerging	= 1 if study focuses on firms in emerging markets	0.184	0.388	0.256
Sector: financial	= 1 if study focuses on financial firms	0.069	0.254	0.099
Estimation techniques				
Method: linear	= 1 if OLS or GLS is used	0.349	0.477	0.376
Method: panel	= 1 if FE or RE is used	0.403	0.491	0.384
Method: IV	= 1 if 2SLS, CF or LIML is used	0.062	0.241	0.055
Method: GMM	= 1 if GMM or its extension is used	0.079	0.270	0.131
Method: other	= 1 if other estimation technique is used (reference category for estimation technique)	0.107	0.309	0.054

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

variables and j indexes the models.

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

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

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

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

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

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

$$Var(\beta_k|D) = \sum_{j=1}^{2^K} \left(Var(\beta_k|D, M_j) + \hat{\beta}_{kj}^2 \right) p(M_j|D) - E(\beta_k|D)^2, \tag{3}$$

A key strength of BMA is its ability to quantify the relative importance of each variable through the posterior inclusion probability (PIP). This statistic sums the posterior model probabilities of all models that include variable X_k, giving the probability that the variable belongs in the true model:

$$PIP_k = \sum_{j\,:\,X_k \in M_j} p(M_j \mid D). \tag{4}$$

Its value ranges from 0 to 1. If PIP = 1, the variable appears in every model with positive posterior probability. Following a scale proposed by Kass and Raftery (1995), the effect of variables may be categorized as weak, positive, strong or decisive, based on PIP value falling into intervals of 0.5–0.75, 0.75–0.95, 0.95–0.99 and 0.99–1 respectively.

With BMA in place, we estimate the following meta-regression:

$$EE_{is} = \beta_0 + \beta_1 SE_{EE_{is}} + \beta_2 X_{is} + \epsilon_{is} \tag{5}$$

where EE_{is} represents the estimated effect, X_{is} is an additional predictor, and $SE_{EE_{is}}$ is standard error. β_0 is a constant, β_1 captures the direction and intensity of publication bias, and ϵ_{is} is the error term.

The meta-regression is estimated using the *bms* package in R, which applies the Markov chain Monte Carlo approach via the Metropolis-Hastings algorithm to avoid the infeasibility of computing all 2^{24} possible models. Instead of evaluating each model, the sampler concentrates on the models with the highest posterior model probabilities, proposing the addition, removal, or swap of regressors and accepting each move with probability proportional to the models' relative marginal likelihoods, weighted by the model prior. Posterior inclusion probabilities are then recovered from the resulting sampled model frequencies (Zeugner, 2011). To ensure that studies contributing many estimates do not dominate the analysis, each observation is weighted by the inverse of the number of estimates per study, giving each study equal total weight. This is consistent with the weighting applied in the frequentist specifications.

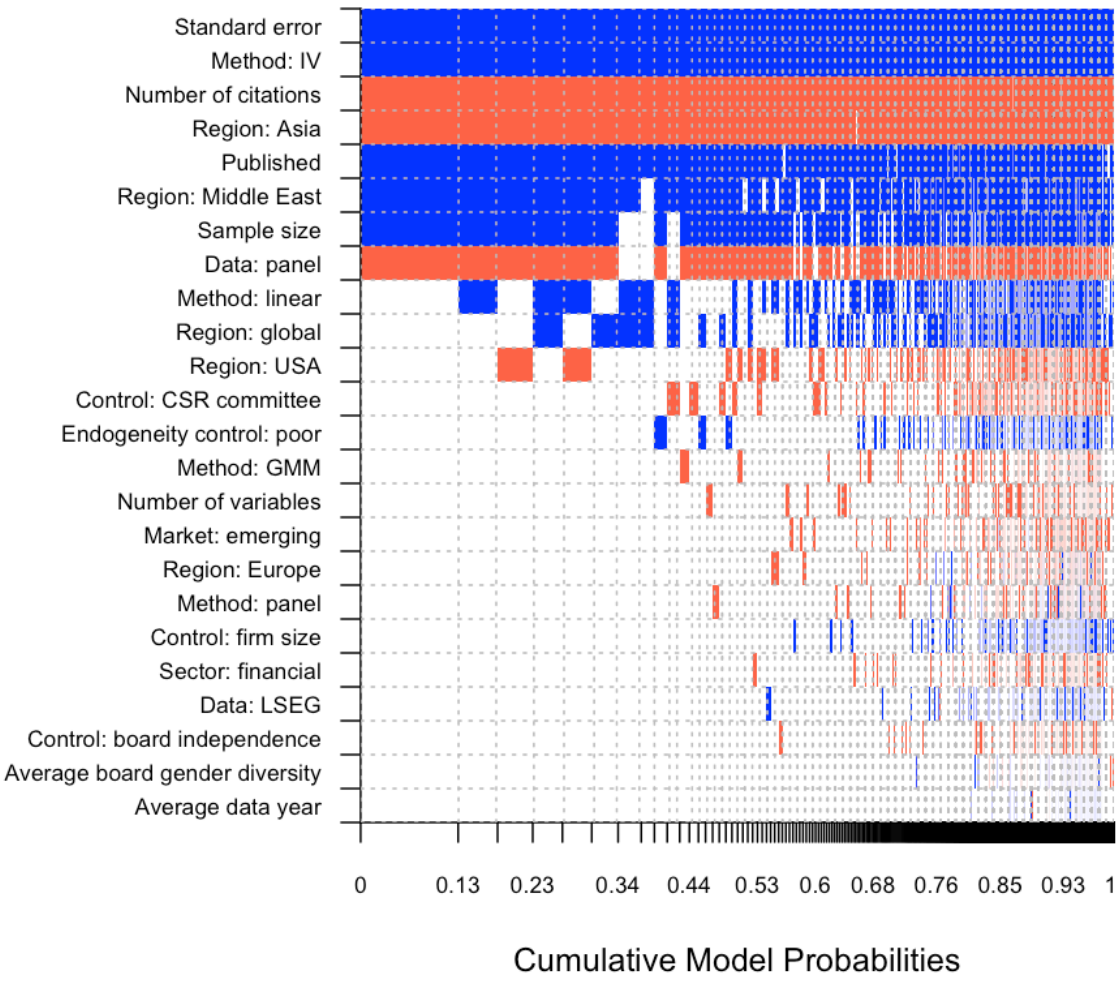
Our baseline BMA estimation adopts the unit information g-prior (UIP), which sets the prior to carry the weight of a single observation. Eicher et al. (2011) recommend it as a robust default for BMA given the limited prior information available to us. To address collinearity, we follow George (2010) and apply a dilution prior. This prior gives less weight to models with a small determinant of the correlation matrix (Hasan et al., 2018). To verify that the sampler reaches its stationary distribution, we report a set of MCMC convergence diagnostics in Appendix C in Table C.1 and Figure C.2.

In addition to the baseline BMA model, we introduce several robustness checks. First, we employ frequentist model averaging (FMA). While BMA incorporates prior beliefs, FMA draws solely from the data at hand, providing a complementary, prior-free check. We follow the meta-analysis implementation of Havránek et al. (2017), who apply the Mallows model averaging estimator of Hansen (2007), whose weights are chosen to minimize the Mallows criterion, an unbiased estimator of the model's squared prediction error that balances in-sample fit against a penalty for the effective number of parameters. This technique, inspired by the work of Magnus et al. (2010) and refined through orthogonalization of the covariate space (Amini and Parmeter, 2012), considerably narrows the model space, which is especially useful when working with a large number of explanatory variables. The implementation builds on openly shared code (Havránek et al., 2024) to facilitate replication. The FMA is estimated over the full set of moderators and is reported alongside the baseline BMA in Table 5, where its coefficients closely track the BMA posterior means and so confirm that the findings do not hinge on the Bayesian priors.

Second, we re-estimate the BMA under an alternative prior structure, the Bayesian risk information criterion (BRIC) g-prior of Fernández et al. (2001) paired with a random model prior. Third, we run a BMA that adopts the same model and g-prior as our baseline BMA, but excludes all Middle East observations. The reason is that the sample board gender diversity is substantially correlated with "Middle East" (the full correlation matrix of the moderators is shown in Figure C.1 in Appendix C), which means that any true dependence of the estimated effect on sample board gender diversity could be masked by collinearity with the Middle East indicator. The results of both are reported in Appendix C (Table C.2).

The results of the baseline BMA estimation are visualized in Figure 4. The vertical axis ranks the explanatory variables based on their posterior inclusion probabilities, while the horizontal axis represents the cumulative posterior model probability. In other words, variables at the top of the figure are the ones that best explain the estimated effects of board gender diversity on ESG scores, and the models on the left offer the optimal balance between goodness of fit and parsimony. Blue color (darker in grayscale) indicates a positive estimated parameter for the corresponding predictor. For example, the color is universally blue for *Middle East*, which suggests that studies focusing on firms in the Middle East typically estimate more positive effects. Red color (lighter in grayscale) indicates a negative estimated parameter. Blank cells mean that the corresponding explanatory variable is excluded from the model. The figure makes it clear that most of the 24 variables do not help explain the systematic differences in reported effect of board gender diversity. Only 8 variables are robustly important, and their corresponding regression coefficients have the same sign irrespective of other controls being included or ignored.

Figure 4: Model inclusion in Bayesian model averaging



Notes: The figure shows results of the baseline BMA estimation reported in Table 5 based on the best 7,359 models (g-prior = UIP, model prior = dilution). The vertical axis ranks the explanatory variables by their posterior inclusion probabilities from highest (top) to lowest (bottom). The horizontal axis represents the cumulative posterior model probability. A blue (darker) color indicates a positive effect, red (lighter) a negative effect, and absence of color denotes non-inclusion of the corresponding variable in a model. Variable descriptions appear in Table 4, and further diagnostic details appear in Appendix C (Table C.1, Figure C.2).

Table 5 reports the quantitative counterpart to Figure 4. In BMA estimations, the marginal effect of a given study characteristic is represented by the posterior mean, while its effectiveness in explaining the differences in the estimated effects reported in the literature is represented by the posterior inclusion probability. For example, when a study employs panel data the estimated board gender diversity premium typically decreases by almost 0.15 points compared to studies that use cross-sectional data, with a posterior inclusion probability of 81%. However, given that employed ESG ratings are measured on a scale of 0 to 100, a change of 0.15 points is modest. Publication characteristics matter

Table 5: Results of baseline BMA and frequentist check estimations

	Bayesian Model Averaging (baseline model)			**Frequentist model averaging (frequentist check)**		
	P. Mean	*P. SD*	*PIP*	*Coef.*	*SE*	*p-value*
Standard error (SE)	2.131	0.086	1.000	2.085	0.095	0.000
Data characteristics						
Sample size	0.030	0.015	0.840	0.027	0.010	0.007
Data: panel	-0.145	0.082	0.808	-0.136	0.052	0.009
ESG data: LSEG	0.001	0.006	0.035	0.024	0.027	0.363
Average data year	0.000	0.002	0.008	0.044	0.020	0.026
Average board gender diversity	0.000	0.003	0.013	-0.020	0.029	0.507
Publication characteristics						
Published	0.156	0.050	0.965	0.152	0.044	0.001
Number of citations	-0.032	0.008	0.994	-0.034	0.007	0.000
Design of the analysis						
Endogeneity control: poor	0.005	0.017	0.106	0.005	0.038	0.898
Number of variables	-0.004	0.015	0.068	-0.062	0.031	0.049
Control variables						
Firm size control	0.004	0.018	0.059	0.090	0.040	0.024
Board independence control	-0.001	0.006	0.034	-0.011	0.026	0.672
CSR committee control	-0.007	0.020	0.151	-0.036	0.027	0.180
Spatial variation						
Region: global	0.036	0.051	0.395	-0.045	0.057	0.430
Region: Europe	-0.003	0.015	0.064	-0.137	0.055	0.013
Region: USA	-0.021	0.040	0.258	-0.190	0.060	0.002
Region: Asia	-0.228	0.046	0.990	-0.210	0.052	0.000
Region: Middle East	0.153	0.076	0.892	0.197	0.078	0.011
Market: emerging	-0.007	0.030	0.067	-0.171	0.060	0.004
Sector: financial	-0.001	0.010	0.047	-0.016	0.041	0.697
Estimation techniques						
Method: linear	0.032	0.039	0.465	-0.012	0.053	0.822
Method: panel	-0.002	0.011	0.061	-0.068	0.052	0.196
Method: IV	0.241	0.051	1.000	0.145	0.069	0.034
Method: GMM	-0.003	0.015	0.069	-0.071	0.065	0.268
Studies	106			106		
Observations	533			533		

Notes: The table displays results for the baseline BMA model and a frequentist model averaging (FMA) estimation based on the Mallows criterion. The FMA is estimated on the full specification and we report its coefficients, standard errors and (two-sided, normal-based) p-values for all moderators. Baseline BMA uses the unit-information g-prior (UIP) and the dilution prior of Eicher et al. (2011) and George (2010). Variables, categories and reference categories are described in Table 4; the omitted reference categories (unpublished papers, proper endogeneity control, other region, other estimation technique) form the respective baselines. P. Mean = Posterior Mean; P. SD = Posterior Standard Deviation; PIP = Posterior Inclusion Probability.

too. Estimated effects from published studies are higher than those from unpublished working papers, and a higher number of citations is associated with a lower board gender diversity premium. The estimation method matters as well, and in fact it carries the largest posterior mean of any moderator: studies that rely on instrumental-variable techniques report premiums about 0.24 points higher, with a posterior inclusion probability

of essentially one. We read this as a methodological rather than a substantive pattern. Instrumental-variable estimates are typically larger and noisier than their OLS or panel counterparts, partly because weak first stages in the primary studies inflate both the coefficient and its standard error, so the higher IV premium more plausibly reflects the estimator than a stronger underlying effect. Yet the largest remaining source of heterogeneity is the geographic location of the firms studied. All BMA specifications, including the baseline and robustness checks, consistently show that studies focusing on Asian firms report a substantially lower board gender premium, while those in the Middle East exhibit a higher one. To understand this result, we examine the common characteristics of the underlying observations. In our dataset, the Asian observations are concentrated mainly in a few Southeast Asian emerging markets, predominantly Malaysia, Indonesia and the Philippines, rather than spread across the continent. These economies are dominated by family business groups and state-linked conglomerates. In such firms, female directors are more likely to be members of the controlling family or appointees added to meet board-diversity rules, while the board itself sits beneath a dominant owner. Either way, the marginal woman director is a weak proxy for the independent, stakeholder-oriented governance that is supposed to raise ESG performance. We therefore read this result as a Southeast Asian emerging-market pattern that heavily overlaps with the *Emerging* indicator. It reflects who serves on these boards and how much power they hold, rather than evidence that board gender diversity matters less in Asia as a whole. In contrast, the studies that cover Middle Eastern countries feature very low average female board representation, typically only a few percent. At first glance, one might expect that such a low initial representation could mechanically amplify the estimated effect of a marginal increase in gender diversity. The data, however, point to a more prosaic reading, consistent with regional selection or unobserved institutional differences rather than a channel this design can identify. From a theoretical standpoint, such small numbers often correspond to tokenism (Kanter, 2008), where female board members lack the critical mass necessary to influence firm strategy or governance, particularly in contexts where gender norms may be restrictive. This raises doubts about the interpretation of the estimated effect as a genuine causal relationship.

Given concerns about the particular interpretation of the *Middle East* category, we conduct a robustness check on a subsample of the original data, excluding observations from this region. As already suggested by the FMA results, other regions may also significantly drive the heterogeneity. The results of this check are presented in Appendix C. Looking at reported inclusion probabilities in Table C.2, we do not observe new predictors and the results remain robust. Our hypothesis that the effect of sample board gender diversity is masked by collinearity with the Middle East indicator is not supported, as the variable's PIP remains below the 0.5 threshold. This reinforces the interpretation that the larger reported effects for the Middle East may reflect other factors. For ex-

ample, it is possible that the few firms in the region appointing women to their boards are not representative. They may already excel in ESG performance due to international ownership, progressive values or external pressure. Alternatively, cultural or publication dynamics may incentivize selective reporting of positive effects.

The analysis of model uncertainty yields two main takeaways. First, publication bias remains robust even with the inclusion of explanatory variables. In fact, the standard-error term remains the clearest signal of publication bias, with posterior inclusion probability equal to one and a large positive coefficient. Second, next to standard error, only seven additional variables out of the remaining 23 cross the PIP threshold of 0.5. The remaining variables included in the analysis do not exhibit a systematic association with the reported effects.

Based on the moderators from our baseline BMA with posterior inclusion probability above 0.5, we construct a "best practice" estimate using the synthetic study approach. The best-practice estimate is the premium that a well-designed study, free of publication bias, would report. We evaluate it separately for each region (Table 6) using a dedicated OLS regression on the high-PIP moderators, distinct from the Mallows frequentist model averaging of Table 5, with the standard error set to zero, a published panel study, and sample size, citations and instrumental-variable use held at their means, varying only the regional indicator. Fuller mechanics are in the table notes. Setting the standard error to zero treats the entire estimate–standard-error correlation as publication selection; because such asymmetry can also reflect between-study heterogeneity (Egger et al., 1997), the best-practice figure is the premium a well-designed study would report under that assumption, not an assumption-free structural effect.

Table 6: Best-practice estimates of the board gender diversity premium by region

	Best-practice estimate	*95% Confidence Interval*
Rest of world (reference)	0.116	(0.080 ; 0.151)
Middle East only	0.276	(0.046 ; 0.506)
Asia only	−0.101	(−0.206 ; 0.004)

Notes: The table reports best-practice estimates of the board gender diversity premium for three mutually exclusive regions. Each estimate evaluates the moderators with posterior inclusion probability above 0.5 at best-practice values, with the standard error set to zero (free of publication bias), a published study using panel data, with the sample size, number of citations and instrumental-variable use held at their sample means, varying only the regional indicator. Estimates and confidence intervals come from a dedicated OLS regression on the high-PIP moderators (distinct from the Mallows frequentist model averaging reported in Table 5), weighted by the inverse number of estimates per study and with standard errors clustered at the study level; the interval is the *lincom* delta-method interval, approximated as the estimate ± 1.96 standard errors.

Table 6 reports the resulting estimates. For the rest of the world, the reference region that excludes the Middle East and Asia, the best-practice premium is 0.116 (95% CI: 0.080 to 0.151). It is small but statistically significant. This is close to, though marginally above, our FAT-PET bias-corrected range of 0.076 to 0.113: once publication bias is removed, the effect a well-designed study would report is modest. On a 0–100 scale, it would be roughly a tenth of an ESG point per one-percentage-point increase in female board representation. The per-percentage-point coefficient, however, understates the practical magnitude of such a change. Take a ten-member board, a common size among listed firms: appointing a single additional woman raises board gender diversity by about ten percentage points. Scaling the bias-corrected estimate linearly over such a change implies a gain of roughly one ESG point, against almost three under the raw, uncorrected literature. Because the relationship is not linear, with the effect larger where women are scarce and smaller where they are already well represented (Figure 1), this is a rough benchmark rather than a precise prediction. A single point is small against the 0–100 scale, but a change of this order is not negligible for a firm close to a rating threshold, and what shrinks under correction is the inflated premium implied by the raw literature, not the corrected relationship itself.

However, the estimates differ sharply across regions. Studies on Middle Eastern firms imply a best-practice premium of 0.276 (95% CI: 0.046 to 0.506). It is about two and a half times the reference and still significantly positive after bias correction. By contrast, studies on Asian firms imply -0.101 (95% CI: -0.206 to 0.004), below the reference and statistically indistinguishable from zero. Geography is thus the dominant source of heterogeneity, with the premium ranging from essentially zero (if anything negative) in Asia, through a small positive effect in most of the world, to a substantially larger one in the Middle East. As discussed above, neither extreme is best read as a genuine difference in how much board gender diversity matters. The Middle East result most plausibly reflects region-specific selection at very low baseline female board representation, while the Asian result is concentrated in a few Southeast Asian emerging markets dominated by family and state-controlled listed firms. The wide Middle East interval, reflecting the small number of Middle Eastern studies, reinforces this caution. Whether the higher Middle Eastern premium reflects a genuine causal effect or residual region-specific heterogeneity remains an open question that warrants dedicated future research.

5 Conclusion

This paper asks whether putting more women on corporate boards reliably improves ESG performance, a question that matters for firms facing gender-equity and sustainability targets at the same time. Drawing on 533 estimates from 106 primary studies, we reach a more cautious answer than the published literature suggests. The raw record implies

that a one-percentage-point increase in female board representation raises ESG scores by about 0.28 points, on average. A positive association remains after correction, and for a board that adds one woman the implied gain is not trivial. But it is much smaller than that headline number, it is concentrated in particular regions, and we cannot rule out that it reflects which firms appoint women rather than what the women themselves do. Correcting for publication bias brings the figure down to between 0.08 and 0.17 points, depending on the method. Our best-practice estimate, which also imposes a sound study design (a panel setup and no publication bias), puts the premium at around 0.12 for most of the world, 0.276 for the Middle East and essentially zero, or even slightly negative, for the Southeast Asian markets that dominate our Asian subsample.

Our analysis yields three main takeaways. First, publication bias is present and robust. The funnel plot is asymmetric and the FAT coefficient stays positive and significant across the main specifications. It survives the inclusion of two dozen control variables and remains of a similar size even when we drop the Middle Eastern studies. Put simply, studies with larger standard errors tend to report larger effects, which is exactly what one would expect when estimates of intuitive sign and statistical significance are easier to publish. The average effect circulating in this literature therefore overstates the effect that survives correction. In the meta-regression, the standard-error term remains the strongest single predictor of the reported effects.

Second, the heterogeneity that remains is systematic and mainly driven by geography. The Middle East enters every BMA specification with a high inclusion probability and a positive sign, while Asia enters with a near-unit probability and a negative one. We do not believe that female directors are simply more effective in the Gulf than in Southeast Asia. A more plausible explanation lies in selection and board composition. In the Middle East, the share of women on boards is very low, often only a few percent, and the few firms that do appoint women tend to be internationally exposed or otherwise unrepresentative, so the estimated effect is inflated by this selection rather than by a stronger underlying channel. The Asian observations, by contrast, come mostly from a few Southeast Asian emerging markets dominated by family and state-linked firms, where a woman on the board is often a member of the controlling family or an appointee added to meet diversity rules, and so a poor proxy for the independent governance that is supposed to lift ESG performance. Both regional results are better understood as composition and selection effects than as evidence on how much board gender diversity really matters.

Third, apart from geography, only a few characteristics systematically shape the reported effects. The estimation method matters most. Studies that rely on instrumental-variable techniques report higher premiums, whereas using panel data lowers the premium by about 0.15 ESG points. Publication status, the number of citations and the sample size also cross the inclusion threshold, while the ESG-rating provider, the control vari-

ables a study uses, the development status of the market and the sector of the firms do not.

A few caveats apply. First, our sample only covers studies that use Bloomberg or LSEG's ESG ratings, which limits comparability with research based on other metrics and, indirectly, tends to restrict the analysis to larger, more visible listed companies, since Bloomberg and LSEG ESG coverage concentrates on such firms. Second, the publication-bias methods we use recover a mean effect conditional on the study-design choices we observe in the literature. They cannot recover a structural causal parameter. The regional patterns described above are exactly the kind of residual heterogeneity that such methods are not designed to resolve. Third, our regional best-practice estimates rest on rather thin evidence at the extremes: the Middle Eastern interval is wide because few studies cover the region, and the Asian estimate, though more precisely pinned down, is barely different from zero.

For policy, board gender diversity remains a goal worth pursuing in its own right, for reasons of fairness, representation, and the quality of governance. Our results do not say that board gender diversity has no effect on ESG. Even after correcting for publication bias, adding one woman to a board of ten raises board gender diversity by about ten percentage points and, as a rough linear benchmark, is associated with a gain on the order of one ESG point; a firm close to a rating threshold may well find a change of this size relevant. What they do question is the much larger payoff implied by the raw literature, since that figure is largely a product of publication bias and of regional and methodological differences rather than of a reliable causal effect. The case for mandating board gender diversity only because it is expected to raise ESG scores is therefore weak. The case for pursuing it on its own merits is not.

References

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

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

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

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

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

S. M. Amini and C. F. Parmeter. Comparison of model averaging techniques:

Assessing growth determinants. *Journal of Applied Econometrics*, 27(5):870–876, 2012.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. Dakhli. Does financial performance moderate the relationship between board attributes and corporate social responsibility in French firms? *Journal of Global Responsibility*, 12(4):373–399, 2021.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. M. Gerged, M. Tran, and E. S. Bed-

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

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

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

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

M. Harjoto, I. Laksmana, and R. Lee. Board Diversity and Corporate Social Responsibility. *Journal of Business Ethics*, 132(4):641–660, December 2015. doi: 10.1007/s10551-014-2343-0.

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

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

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

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

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

T. Havránek, T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, and R. C. van Aert. Reporting Guidelines for Meta-Analysis in Economics. *Journal of Economic Surveys*, 34(3):469–475, 2020.

T. Havránek, Z. Iršová, L. Laslopová, and O. Zeynalova. Publication and Attenuation Biases in Measuring Skill Substitution. *The Review of Economics and Statistics*, 106(5):1187–1200, 2024.

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

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

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

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

Z. Iršová, H. Doucouliagos, T. Havránek, and T. Stanley. Meta-analysis of social science research: A practitioner's guide. *Journal of Economic Surveys*, 38

(5):1547–1566, 2024.

Z. Iršová, P. R. D. Bom, T. Havránek, and H. Rachinger. Spurious precision in meta-analysis of observational research. *Nature Communications*, 16:8454, 2025.

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

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

M. Kamran, H. G. Djajadikerta, S. Mat Roni, E. Xiang, and P. Butt. Board gender diversity and corporate social responsibility in an international setting. *Journal of Accounting in Emerging Economies*, 13(2):240–275, 2023.

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

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

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

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

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

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

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

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

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

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

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

B. Miranda, C. Delgado, and M. C. Branco. Board characteristics, social trust and

ESG performance in the European banking sector. *Journal of Risk and Financial Management*, 16(4):244, 2023.

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

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

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

C. Post and K. Byron. Women on Boards and Firm Financial Performance: A Meta-Analysis. *Academy of Management Journal*, 58(5):1546–1571, 2015.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

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

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

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

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

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

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

T. D. Stanley. 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, 2008.

T. D. Stanley and H. Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2014.

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

T. D. Stanley, H. Doucouliagos, M. Giles, J. H. Heckemeyer, R. J. Johnston, P. Laroche, J. P. Nelson, M. Paldam, J. Poot, G. Pugh, R. S. Rosenberger, and K. Rost. Meta-analysis of economics re-

search reporting guidelines. *Journal of Economic Surveys*, 27(2):390–394, 2013.

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

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

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

R. C. van Aert and M. A. van Assen. Correcting for publication bias in a meta-analysis with the p-uniform* method. *Psychonomic Bulletin & Review*, 33(3): 102, 2026.

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

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

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

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

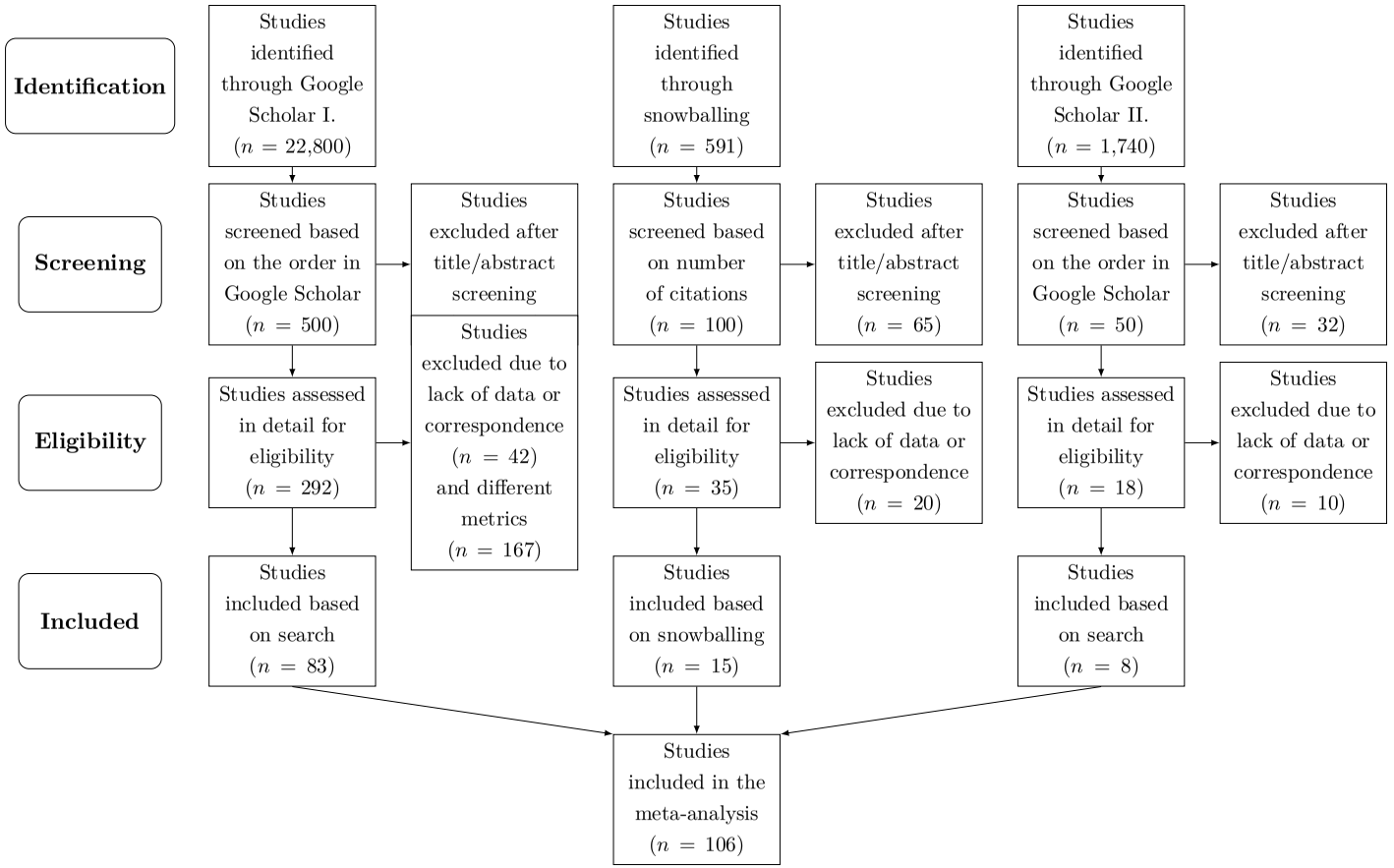
Q. Wu, F. Furuoka, and S. C. Lau. Corporate social responsibility and board gender diversity: a meta-analysis. *Management Research Review*, 45(7):956–983, 2022.

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

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

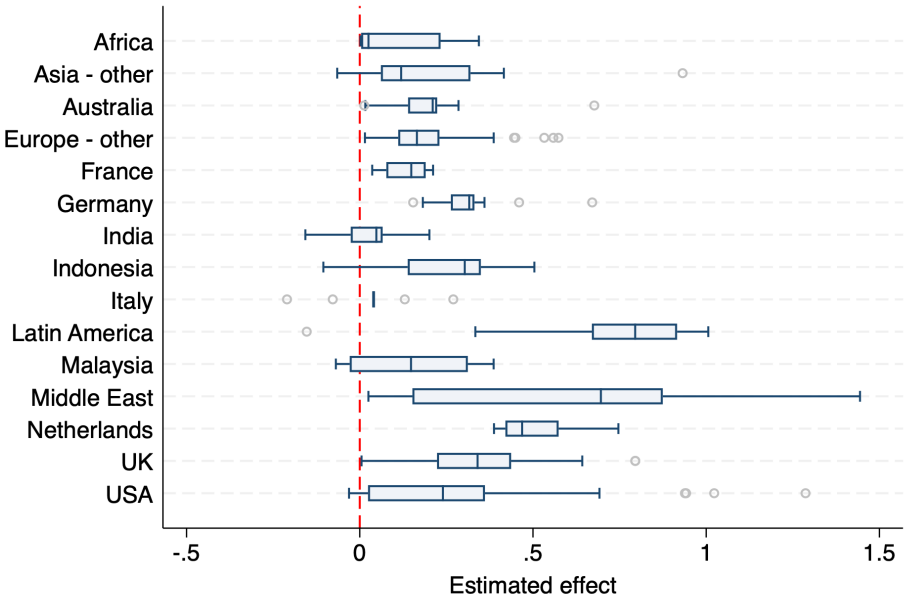
Appendix A

Figure A.1: PRISMA flow diagram



Notes: The figure presents the PRISMA flow diagram for the study-selection procedure (Iršová et al., 2024; Page et al., 2021; Stanley et al., 2013). Studies are identified in Google Scholar because it searches the full text of articles rather than just titles, abstracts, and keywords. The following query is used: **"Board" AND "Gender" AND "Diversity" AND "ESG" AND ("Disclosure" OR "Score" OR "Performance")**. The initial search in May 2024 returned more than 22,800 records, of which we screen the first 500 by relevance ranking. Backward snowballing is then performed by compiling the 100 most frequently cited works among the included studies. The search was closed on December 16, 2024, when we ran a further Google Scholar query restricted to 2024 to capture articles published between May and December. Whenever a study had a plausible chance of containing usable estimates, its full text was read and assessed. The complete dataset is available in the online appendix at meta-analysis.cz/esg.

Figure A.2: Variation of estimates within and across countries



Notes: The figure shows a box plot of the estimated effects within the interquartile range. The inner box line indicates the median. The whiskers extend to the most extreme values within 1.5 times the interquartile range.

Appendix B

Table B.1: Robustness checks: excluding the Middle East and unwinsorized data

Block 1: Subset excluding Middle East					
Panel A: Linear	OLS	Between Effects	Study Weight	Precision Weight	MAIVE
Publication Bias	1.689***	2.128***	1.628***	2.157***	
(*Standard Error*)	(0.325)	(0.218)	(0.347)	(0.324)	
	[0.853, 2.413]		[0.622, 2.328]	[1.410, 2.859]	
Mean Beyond Bias	0.116***	0.071**	0.120***	0.080***	-0.062
(*Constant*)	(0.022)	(0.031)	(0.029)	(0.020)	(0.168)
	[0.072, 0.162]		[0.055, 0.183]	[0.033, 0.125]	{-0.239, 0.190}
First-stage robust F-stat					3.13
Panel B: Nonlinear	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		$P = 0.281$		1.794**	
		(0.046)		(0.887)	
Effect Beyond Bias	0.087***	0.114***	0.166***	0.084***	0.167***
	(0.006)	(0.010)	(0.013)	(0.004)	(0.027)
# of estimates	218		100		
% of information			100%		
Observations	503	503	503	503	503
Studies	100	100	100	100	100
Block 2: Unwinsorized data					
Panel A: Linear	OLS	Between Effects	Study Weight	Precision Weight	MAIVE
Publication Bias	0.239	0.376***	0.196	1.527***	
(*Standard Error*)	(0.117)	(0.054)	(0.074)	(0.448)	
	[-9.502, 2.109]		[-9.880, 1.940]	[0.554, 2.620]	
Mean Beyond Bias	0.248***	0.246***	0.307***	0.083***	0.089
(*Constant*)	(0.032)	(0.053)	(0.050)	(0.019)	(0.067)
	[0.183, 0.317]		[0.204, 0.417]	[0.037, 0.125]	{-0.005, 0.367}
First-stage robust F-stat					0.97
Panel B: Nonlinear	WAAP	Selection Model	STEM method	Endogenous Kink	p-uniform*
Publication Bias		$P = 0.277$		3.821***	
		(0.042)		(1.103)	
Effect Beyond Bias	0.038***	0.113***	0.172***	0.060***	0.159***
	(0.005)	(0.010)	(0.013)	(0.003)	(0.038)
# of estimates	178		103		
% of information			99.9%		
Observations	533	533	533	533	533
Studies	106	106	106	106	106

Notes: Block 1 excludes Middle East estimates; Block 2 uses unwinsorized data. *Panel A:* FAT-PET regression $E_{is} = \beta_0 + \beta_1 \cdot \text{SE}(E_{is}) + \varepsilon_{is}$; where β_1 is publication bias and the constant β_0 is the mean beyond bias. Cluster-robust standard errors are in parentheses, wild-bootstrap CIs in square brackets (Roodman et al., 2019) and the Anderson and Rubin (1949) 95% CI in curly brackets for MAIVE by Iršová et al. (2025) that corrects for spurious precision. *Panel B:* WAAP denotes weighted average of adequately powered estimates (Stanley et al., 2017); Selection model denotes the technique due to Andrews and Kasy (2019); Stem denotes the stem-based technique (Furukawa, 2019); Kink denotes the endogenous kink model (Bom and Rachinger, 2019); p-uniform* denotes the technique due to van Aert and van Assen (2026). Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

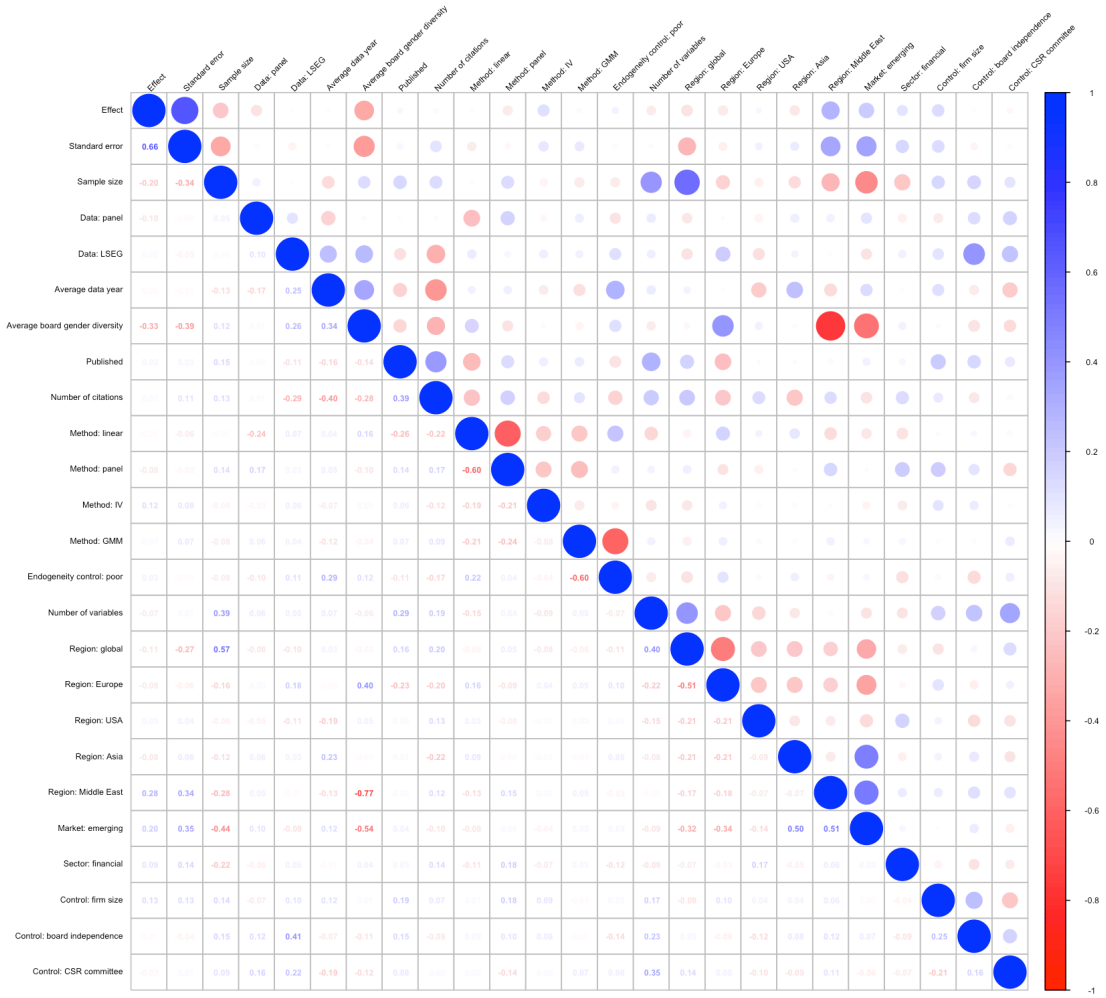
Table B.2: Provider invariance: interaction FAT-PET test

	OLS	Precision Weight
Publication bias (β_1)	1.624***	2.350***
	(0.304)	(0.392)
× Bloomberg (β_3)	0.397	-0.413
(difference)	(0.610)	(0.636)
	[-1.206, 1.783]	[-2.041, 0.988]
Mean beyond bias (β_0)	0.135***	0.075***
	(0.026)	(0.023)
Bloomberg (β_2)	-0.057	0.011
(difference)	(0.042)	(0.039)
	[-0.149, 0.037]	[-0.113, 0.086]
Observations	533	533
Studies	106	106

Notes: Estimates from the interacted FAT-PET regression $E_{is} = \beta_0 + \beta_1 \,\mathrm{SE}(E_{is}) + \beta_2 \,\mathrm{Bloomberg}_s + \beta_3 \,[\mathrm{SE}(E_{is}) \times \mathrm{Bloomberg}_s] + \varepsilon_{is}$, where $\mathrm{Bloomberg}_s = 1$ if the rating provider is Bloomberg (baseline = LSEG ESG data provider). β_1 and β_0 are the funnel-asymmetry slope (publication bias) and corrected mean for the baseline provider; β_3 tests whether publication bias differs across providers and β_2 whether the corrected mean differs. Cluster-robust standard errors (clustered by study) in parentheses. For the difference terms (β_3, β_2), 95% wild-bootstrap confidence intervals (Roodman et al., 2019) are in square brackets and significance is from wild-bootstrap p-values. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

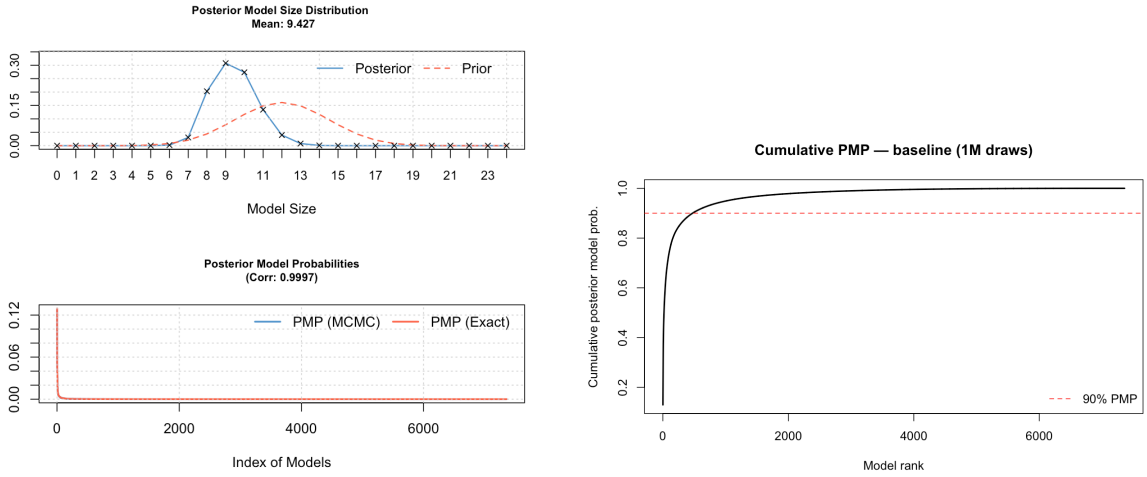
Appendix C

Figure C.1: Correlations between potential predictors of heterogeneity



Notes: The figure displays correlation coefficients for variables used in heterogeneity analysis.

Figure C.2: Model size and convergence for the baseline BMA



(a) Posterior model size and model probabilities (b) Cumulative posterior model probability

Notes: Panel (a) depicts the posterior model size distribution and the posterior model probabilities; panel (b) shows the cumulative posterior model probability across the best models, with the dashed line marking the 90% threshold. Both refer to the baseline BMA estimation reported in Table 5, using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.

Table C.1: Diagnostics of the baseline BMA

Sampler & model space		**Convergence & priors**	
MCMC draws	$1 \cdot 10^6$	Corr. PMP	0.9997
Burn-in draws	$3 \cdot 10^5$	Top models (% PMP)	100%
Run time	6.98 mins	Shrinkage (avg.)	0.9981
Mean no. regressors	9.4270	Model prior	Dilution / 12
Models visited	183,280	g-prior	UIP
Model space (2^K)	$1.7 \cdot 10^7$	No. observations	533
of which visited	1.1%		

Notes: The table shows technical diagnostic information of baseline BMA estimation reported in Table 5, using the unit information g-prior (UIP) and the dilution model prior due to Eicher et al. (2011) and George (2010), respectively.

Table C.2: Baseline BMA robustness: alternative priors and Middle East exclusion

	Prior robustness *(BRIC g-prior, random model prior)*			**Subsample robustness** *(UIP, dilution; excl. Middle East)*		
	P. Mean	*P. SD*	*PIP*	*P. Mean*	*P. SD*	*PIP*
Standard error (SE)	2.134	0.087	1.000	2.105	0.089	1.000
Data characteristics						
Sample size	0.035	0.011	0.960	0.031	0.013	0.897
Data: panel	-0.165	0.063	0.936	-0.170	0.071	0.901
ESG data: LSEG	0.001	0.006	0.039	0.003	0.012	0.078
Average data year	0.001	0.005	0.058	0.000	0.001	0.006
Average board gender diversity	0.000	0.005	0.041	0.000	0.002	0.011
Publication characteristics						
Published	0.158	0.047	0.981	0.161	0.048	0.976
Number of citations	-0.032	0.007	0.997	-0.031	0.007	0.992
Design of the analysis						
Endogeneity control: poor	0.005	0.017	0.114	0.009	0.023	0.168
Number of variables	-0.014	0.028	0.247	-0.003	0.013	0.055
Control variables						
Firm size control	0.007	0.024	0.108	0.004	0.018	0.062
Board independence control	-0.001	0.006	0.035	-0.001	0.006	0.035
CSR committee control	-0.004	0.015	0.099	-0.002	0.011	0.074
Spatial variation						
Region: global	0.023	0.041	0.295	0.027	0.044	0.326
Region: Europe	-0.004	0.020	0.086	-0.002	0.012	0.053
Region: USA	-0.022	0.043	0.264	-0.021	0.039	0.280
Region: Asia	-0.228	0.043	0.997	-0.225	0.042	0.994
Region: Middle East	0.161	0.075	0.903	— (excluded) —		
Market: emerging	-0.009	0.033	0.110	-0.007	0.028	0.073
Sector: financial	-0.001	0.008	0.033	-0.002	0.013	0.062
Estimation techniques						
Method: linear	0.022	0.033	0.361	0.026	0.036	0.404
Method: panel	-0.002	0.011	0.061	-0.003	0.014	0.083
Method: IV	0.233	0.050	1.000	0.261	0.050	1.000
Method: GMM	-0.003	0.014	0.058	-0.002	0.013	0.059
Studies	106			100		
Observations	533			503		

Notes: The table reports two robustness checks on the baseline BMA (Table 5). The *prior robustness* check (left) re-estimates the full sample with the BRIC g-prior (Fernández et al., 2001) and the random model prior (Ley and Steel, 2009). The *subsample robustness* check (right) retains the unit-information g-prior (UIP) and the dilution model prior of Eicher et al. (2011) and George (2010) but excludes Middle East observations, so this regressor drops out of the specification. Both checks reproduce the same set of moderators that cross the PIP > 0.5 threshold in the baseline; the frequentist check is reported once, for the baseline, in Table 5. Variables are described in Table 4. P. Mean = Posterior Mean; P. SD = Posterior Standard Deviation; PIP = Posterior Inclusion Probability.