



UvA-DARE (Digital Academic Repository)

The business case for demographic diversity in strategic leadership teams

A systematic and critical review of the causal evidence

Sieweke, Jost; Hentschel, Tanja; Gazdag, Brooke A.; Henningsen, Levke

DOI

[10.1016/j.leaqua.2024.101843](https://doi.org/10.1016/j.leaqua.2024.101843)

Publication date

2025

Document Version

Final published version

Published in

Leadership Quarterly

License

CC BY-NC

[Link to publication](#)

Citation for published version (APA):

Sieweke, J., Hentschel, T., Gazdag, B. A., & Henningsen, L. (2025). The business case for demographic diversity in strategic leadership teams: A systematic and critical review of the causal evidence. *Leadership Quarterly*, 36(1), Article 101843. <https://doi.org/10.1016/j.leaqua.2024.101843>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.



Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

The business case for demographic diversity in strategic leadership teams: A systematic and critical review of the causal evidence

Jost Sieweke^a, Tanja Hentschel^b, Brooke A. Gazdag^{c,*}, Levke Henningsen^d

^a Vrije Universiteit Amsterdam, School of Business and Economics, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands

^b University of Amsterdam, Amsterdam Business School, Postbus 15953, 1001 NL Amsterdam, the Netherlands

^c Kühne Logistics University (KLU), Grosser Grasbrook 17, 20457 Hamburg, Germany

^d University of Exeter, Rennes Drive, Exeter EX4 4PU, United Kingdom

ARTICLE INFO

Keywords:

Business case for diversity
TMT gender diversity
Board gender diversity
Quasi-Experiments
Endogeneity

ABSTRACT

Demographic diversity (e.g., gender, age, race, ethnicity) in strategic leadership teams (i.e., boards of directors and top management teams) has received global attention recently. Policymakers have promoted diversity policies by citing the “business case” for diversity that suggests a positive (causal) effect on firm performance. Our focus is twofold: First, we systematically evaluate the methodological rigor of 64 studies on the relationship between strategic leadership team demographic diversity and firm performance (1994–2023) from Financial Times (FT) 50 journals, finding that ca. 70 percent show implausible causal effects, ca. 20 percent lack sufficient information, and only 11 percent (N = 7) demonstrate plausible causal effects. Second, we synthesize research findings of the seven studies. The five studies on gender diversity yield mixed results: some report positive or negative effects, whereas the majority finds no effects on firm performance. Regarding ancestral and genetic diversity, the studies support the business case argument. Overall, our review provides three key insights: (1) a critical evaluation of the causal evidence regarding the business case for demographic diversity in strategic leadership teams, (2) a synthesis of the research findings by focusing on rigorously conducted studies, and (3) hands-on recommendations for refining future approaches for causal research.

Introduction

In the last 20 years, the issue of demographic diversity within strategic leadership teams – that is, boards of directors and top management teams (TMT) – has gained significant attention on the global stage. To that end, legislation was introduced to direct organizations to increase diversity (Ben-Shahar et al., 2024). For example, countries such as Norway and France, and U.S. states such as California have introduced quotas (e.g., gender quotas) in boards and top management teams that aim at increasing demographic diversity. A key argument used by policy-makers to justify the introduction of strategies and initiatives to establish diverse leadership teams was the so-called “business case” for

diversity (e.g., Herring, 2009; Hoobler et al., 2018; Roberson et al., 2017), which refers to the supposed positive influence of demographic diversity on firm performance. To justify their position, policy-makers also refer to research providing evidence for the business case.¹

Whereas it is instrumental that policy-makers use research evidence to justify decisions, it is unclear whether the methodological quality of the studies on strategic leadership team demographic diversity is sufficient for informing policy. As Antonakis (2017, p. 9) argued, “to influence policy and practice, it is causal effects that we should go after.” Yet, identifying the causal effect of strategic leadership team demographic diversity on firm performance is difficult, especially in field settings, because of endogeneity concerns. Endogeneity refers to a situation in a

* Corresponding author.

E-mail addresses: j.sieweke@vu.nl (J. Sieweke), t.hentschel@uva.nl (T. Hentschel), brooke.gazdag@klu.org (B.A. Gazdag), l.henningsen@exeter.ac.uk (L. Henningsen).

¹ For instance, the state of California argued in its legislation that required publicly traded companies to have a minimum number of female directors on their boards (State of California, 2018) that this policy “will boost the California economy” (Section 1a).

<https://doi.org/10.1016/j.leaqua.2024.101843>

Received 15 September 2023; Received in revised form 18 October 2024; Accepted 21 October 2024

Available online 16 November 2024

1048-9843/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

statistical model where an explanatory variable is correlated with the error term, potentially leading to biased and inconsistent estimates. There are different sources of endogeneity in research on strategic leadership team demographic diversity of which omitted variable bias and simultaneity are particularly relevant (Sieweke et al., 2023; Yang et al., 2019). First, *omitted variable bias* is when one or more relevant variables (i.e., variables correlated with both the independent and dependent variables) are excluded from the model. For example, a firm characteristic such as organizational culture may affect women's promotion to top management positions and firm performance. Omitting this variable from the model will bias the estimates. Second, *simultaneity* indicates that the independent and dependent variables influence each other. For instance, higher firm performance may positively influence a greater board demographic diversity, and a greater board diversity will subsequently increase firm performance. Importantly, endogeneity can bias findings "to the point where results cannot inform policy" (Antonakis, 2017, p. 14). As such, to ensure that policies using the business case are effective, endogeneity concerns must be addressed. To address these endogeneity concerns, it is crucial to analyze the quality of the causal evidence regarding the relationship between strategic leadership team demographic diversity and firm performance.

Because policy-makers rely on "business case" arguments for demographic diversity in strategic leadership teams, our paper aims to understand better and closely scrutinize the scientific quality of the causal evidence on the business case for strategic leadership team demographic diversity. To do so, we conduct a systematic review of research on the relationship between strategic leadership team demographic diversity and firm performance.² Our review lies at the intersection of methodological reviews, which focus on a specific methodological issue (Aguinis et al., 2023), and integrative reviews, which synthesize the knowledge of a topic (Cronin & George, 2023). We systematically collect and analyze studies focusing on the relationship between strategic leadership team demographic diversity and firm performance; yet, instead of synthesizing the research findings of all studies, we first carefully analyze the methodological rigor of these studies regarding their identification of causal effects. For instance, we not only check whether a study uses an instrumental variable (IV) approach to identify causal effects between strategic leadership team demographic diversity and firm performance but also whether the authors follow common guidelines in the IV literature (e.g., Bastardo et al., 2023) to assess the overall quality of the causal evidence. With this approach, we aim to distinguish substantial from more symbolic applications of identification strategies. In the next step, we synthesize the research findings of those studies that use identification strategies substantially to provide an overview of the current state of plausible causal evidence regarding the relationship between strategic leadership team demographic diversity and firm performance.³

Our review makes three contributions: (1) to evaluate the quality of the available *causal* evidence for the "business case" for demographic diversity in strategic leadership teams, (2) to synthesize the evidence of

² In academic research, the topic of demographic diversity in TMTs and boards has received much attention spanning several disciplines—including leadership, management, finance, and economics. However, the conversations about TMTs and boards often occur in parallel research streams and are seldomly integrated (for an exception, see Roberson et al., 2017) despite both their individual and joint influence on firm outcomes (Luciano et al., 2020). To assess the evidence for the business case for diversity, we review and evaluate the literature on both TMT and board demographic diversity, which we refer to as strategic leadership team demographic diversity in this review.

³ We only focus on the causal evidence from a statistical perspective and explicitly do not analyze and evaluate the quality of the applied theories. Despite the importance of theories explaining why a statistical relationship can be observed, we believe rigorously identifying causal relationships is valuable, especially for policy-makers (Antonakis, 2017). We discuss this aspect in more detail in our "Recommendations for Future Research" section.

rigorous studies testing the "business case" for diversity, and (3) to provide hands-on recommendations for how to conduct such research with appropriate methodological rigor. Furthermore, we regard this literature review as a timely contribution to the broader practical leadership field and policy formation: Due to the reliance on "business case" arguments for demographic diversity, our review helps to inform policy-makers and business leaders to make sound decisions regarding their diversity policies for strategic leadership teams (Antonakis, 2017).

Background

Our primary research question asks about the quality and reliability of the *causal* evidence for the business case of demographic diversity in strategic leadership teams. To answer this question, the current manuscript is organized into two major sections. First, we focus on the research designs employed in studies on demographic diversity in strategic leadership teams (i.e., on boards and in top management teams) and firm performance (i.e., financial as well as social and environmental). Specifically, we consider the following demographic diversity categories: (1) ethnicity/race/minority, (2) age, (3) nationality/foreigner/cultural, and (4) gender. Taking an additional step compared to other reviews, we conduct an in-depth evaluation of the quality of the studies' research designs. That is, we investigate whether the research designs' assumptions for causal identification are tested and met. We analyzed the respective literature for each research design to identify (a) key assumptions and (b) statistical tests to check for these assumptions (e.g., Bastardo et al., 2023; Hausman & Rapson, 2018; Narita et al., 2023). Once we determined that studies both relied on a causal design and applied it accurately and precisely, we summarized the trends in this study subsample. By summarizing only the most rigorous studies, our review goes over and above meta-analyses (and other reviews) that give equal weight to all studies independent of their methodological rigor. Second, based on this review, we provide a future research agenda for how to improve future research, including a primer on identification strategies, a decision tree for selecting an appropriate strategy, and a checklist for researchers to ensure the methodological rigor of their research on strategic leadership team diversity and firm performance – ultimately to inform policy.

Methods

Search

This article focuses on the causal relationship between strategic leadership team demographic diversity and firm performance. Given that we are interested only in the *causal* relationship, we systematically searched top journals because we expect studies published in these journals to be of a greater methodological quality, at least on average. Therefore, we followed prior systematic literature reviews in the field of leadership (e.g., Krause et al., 2022; Van Doorn et al., 2023) and searched journals included in the Financial Times (FT) 50 journal list. The FT50 list includes journals from multiple business-related disciplines (e.g., management, information systems, accounting, finance) and economics, which is essential given that strategic leadership teams and firm performance are relevant topics in all these disciplines. In addition to the FT50 journals, we added five journals because they focus on leadership-related (The Leadership Quarterly), corporate governance-related (Corporate Governance: An International Review, Corporate Governance – The International Journal of Business in Society), or non-financial firm performance topics (Corporate Social Responsibility and Environmental Management) and because of their status as general interest business journal (Journal of Business Research). We searched the title, abstract, and author keywords of all articles published between 1994–2023 using a list of search terms

referring to strategic leadership teams, demographic diversity, and different types of firm performance.⁴

Our search identified a sample of 220 articles published between the years 1994 and 2023. In the next step, we read the titles and abstracts of these articles to determine whether they should be included in the full-text analysis. At this stage, we excluded 130 articles based on five exclusion criteria: no focus on demographic diversity ($n = 74$), strategic leadership teams ($n = 27$), or firm performance ($n = 16$); no use of demographic diversity as independent variable ($n = 2$); and a lack of (primary) quantitative data ($n = 11$).⁵ We then read the full text of the remaining 90 articles and excluded 26 articles for various reasons: no focus on firm performance ($n = 10$), no focus on demographic diversity ($n = 9$), demographic diversity was only a moderator ($n = 3$), demographic diversity was combined with task-related diversity ($n = 1$), no focus on strategic leadership teams ($n = 1$), machine-learning approach ($n = 1$), and a qualitative methodology ($n = 1$). The final sample of our literature review was 64 articles. An overview of all articles can be found in Appendix A. Fig. 1 shows a PRISMA flow diagram, which describes our search process.

Coding

The coding of the final sample of 64 articles followed a two-step procedure: In the first step, the first author coded all method-related information. More precisely, we focused on the applied identification strategies, which are “strategies for attempting to draw causal inference from observational data” (Athey & Imbens, 2017, p. 4). For the coding of the identification strategies, we built on the work of Antonakis et al. (2010) and Angrist and Pischke (2009), who both provide a comprehensive overview of identification strategies. Specifically, we used the following codes: statistical adjustment, matching, instrumental variable (IV) design, difference-in-differences (DID) design, regression discontinuity (RD) design, Heckman selection, and difference and system Generalized Method of Moments (GMM). We coded all identification strategies used in an article and did not differentiate whether a strategy was used as a main or additional analysis.

In the second step, we coded the plausibility of the article reporting causal estimates of the relationship between strategic leadership team demographic diversity and firm performance. This plausibility analysis considers that each identification strategy relies on assumptions for causal inference. Not meeting these assumptions limits the extent to which a causal effect can be identified; in some cases, estimates can be even more biased than simply using ordinary least squares (OLS) estimates. For instance, simulations by Semadeni and colleagues (2014) show that an endogenous instrument produces severely biased estimates that are inferior to OLS estimates; the same pattern also applies to DID estimates if the parallel trend assumption is violated (O’Neill et al., 2016). Therefore, we coded all information referring to the key assumptions of each identification strategy, which we show in Table 1. We

⁴ Specifically, we used the following keyword combination: (“top management” OR “top management team*” OR “TMT” OR “executive*” OR “executive team*” OR “strategic leader*” OR “board* of director*” OR “boardroom” OR “board*”) AND (“diversit*” OR “heterogene*” OR “gender” OR “sex” OR “male” OR “female” OR “wom?n” OR “m?n” OR “ethnic*” OR “race” OR “minorit*” OR “age” OR “old*” OR “young*” OR “nationalit*” OR “foreign*” OR “cultur*”) AND (“firm perform*” OR “firm effic*” OR “firm profit*” OR “firm valu*” OR “compan* perform*” OR “compan* effic*” OR “compan* profit*” OR “company valu*” OR “corporate perform*” OR “corporate effic*” OR “corporate profit*” OR “corporate valu*”). Readers can replicate our search using this link: <https://www.webofscience.com/wos/woscc/summary/112b8d7e-4fe1-4884-8439-edc2ed8db7c8-df6946a4/relevance/1>.

⁵ The first author was responsible for this part of the selection process. The fourth author independently coded 50 percent ($n = 110$) of the articles. The two coders initially agreed on 106 out of 110 articles (96.4 percent) and resolved their disagreement after checking the articles’ full text.

identified the key assumptions by analyzing several key sources, which we also listed in Table 1.

Based on the information provided in the articles, we classified the studies into three categories: plausible causal effect, implausible causal effect, and insufficient information. We classified an article in the “plausible causal effect” category if (a) it reports information regarding all key assumptions of the identification strategy and (b) the information indicates that the assumptions are probably not violated. An article was classified in the “implausible causal effect” category if (a) the article only used statistical adjustment⁶ or (b) the information indicates that at least one of the key assumptions is probably violated. Finally, an article was classified in the “insufficient information” category if at least one piece of information regarding a key assumption of the identification strategy is not reported. Only seven articles (10.9 percent) were classified as “plausible causal effect,” whereas 44 articles (68.8 percent) were classified as “implausible causal effect” and 13 (20.3 percent) as providing “insufficient information” for a thorough assessment.

Results

Part I: Review of the identification strategies

We analyze the identification strategies used to test the causal relationship between strategic leadership team diversity and firm performance. This analysis is a crucial step because the quality of the design affects the extent to which researchers and practitioners (e.g., executives and policy-makers) can build on the insights from prior research. Our findings, summarized in Table 2, provide an overview of the identification strategies. It is important to note that studies can use multiple identification strategies. The table shows that statistical adjustment is the most frequently applied identification strategy (93.8 percent), followed by the instrumental variable design (25.0 percent), and difference and system Generalized Method of Moments (17.2 percent). Only a few articles use matching methods (3.1 percent) and the difference-in-differences (DID) design (1.6 percent); no article applied the regression discontinuity design (RDD) or the Heckman selection. Below, we briefly introduce the identification strategies (i.e., statistical adjustment, the instrumental variable design, difference and system generalized method of moments (GMM), matching design, and difference-in-difference design) and explain the key assumptions for valid causal inference. We then discuss to what extent the articles in our sample meet these assumptions and whether they can be classified as providing “plausible causal evidence.”

Statistical adjustment

Statistical adjustment refers to the idea of identifying a treatment’s causal effect by including all relevant confounders (Antonakis et al., 2010). We find that 93.8 percent of the articles (i.e., 60 articles) use statistical adjustment as an identification strategy.

Studies using statistical adjustment try to identify causal effects by including a variety of control variables at the strategic leadership team (e.g., size), firm (e.g., size, industry), and industry level (e.g., dynamism). The control variables are often time-varying, meaning they change over time. In addition to these time-varying covariates, most studies seek to reduce the potential for omitted variable bias by controlling for time-invariant firm or industry characteristics and firm – or industry-invariant period effects by adding fixed effects.

Despite the frequent use of statistical adjustment as an identification strategy, this strategy has some severe limitations. Most importantly, causal identification with statistical adjustment depends on the unconfoundedness assumption; we must include *all* relevant confounding

⁶ We assume it is (almost) impossible for researchers to control for all time-invariant and time-varying cofounders, which is why an omitted variable bias is highly likely in the case of statistical adjustment.

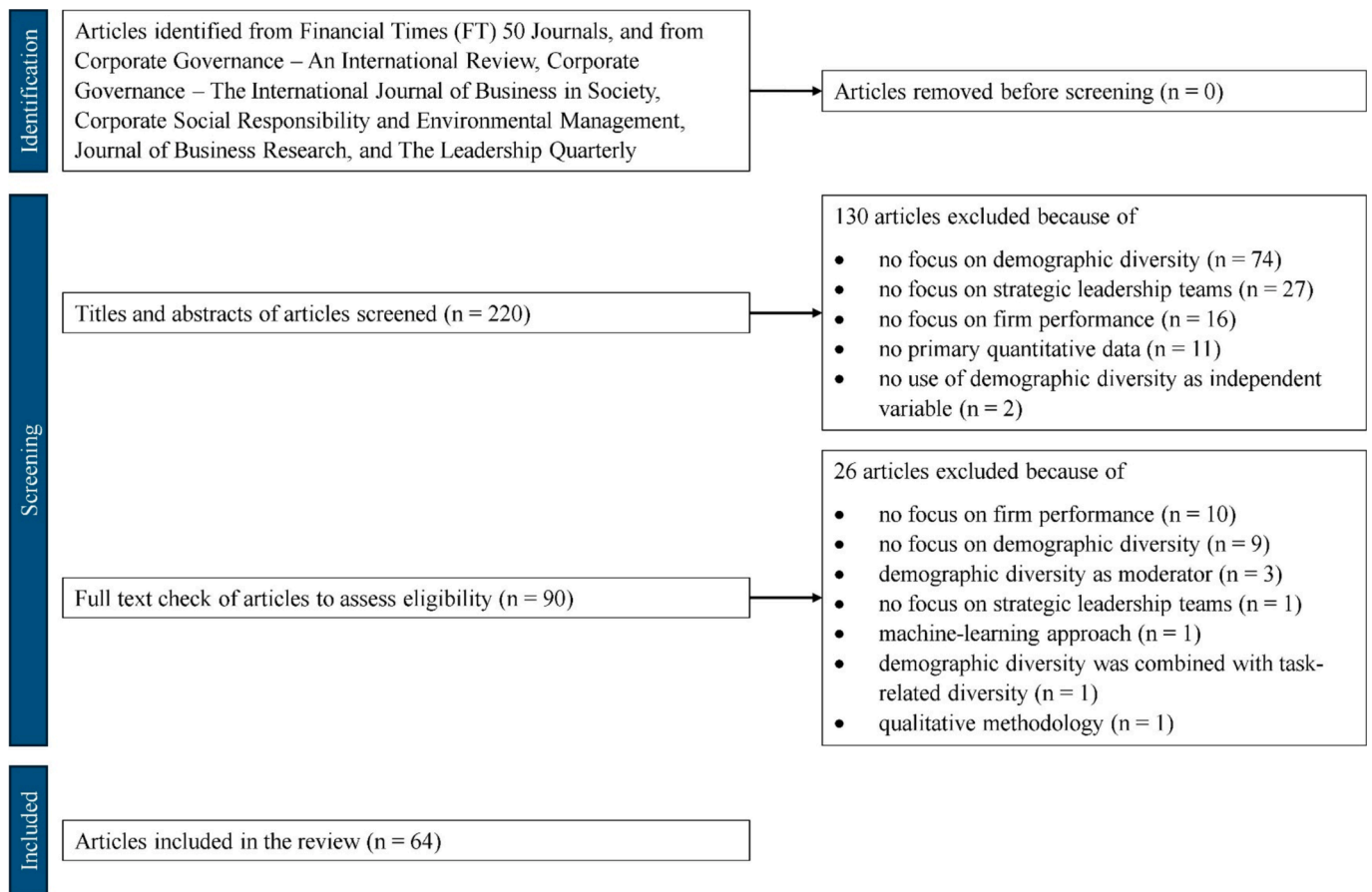


Fig. 1. PRISMA Flow Diagram.

variables in our analysis (Angrist & Pischke, 2009). This assumption is likely to be violated because multiple (and mostly unobserved) variables may affect the relationship between strategic leadership team demographic diversity and firm performance (e.g., Sieweke et al., 2023; Yang et al., 2019). Even controlling for fixed effects (e.g., firm, industry, time) effects does not alleviate the concerns. Although this approach reduces the number of potential confounders, we should not assume that it identifies causal effects because (a) not all confounders are time-invariant, and (b) we cannot control for all time-varying confounders (Hill et al., 2020). Therefore, we classified all articles using statistical adjustment as an identification strategy—even when including (two-way) fixed effects—as “implausible causal effect” because of the likely omission of confounders.

The instrumental variable design (IV Design)

The IV design seeks to identify the causal effect of a treatment by using an exogenous variable – the instrument – that is strongly related to the treatment but otherwise unrelated to the dependent variable. Sixteen articles (25 percent) apply the IV design to estimate the causal effect of strategic leadership team demographic diversity on firm performance.

Causal inference using the IV design builds on several assumptions, most importantly instrument exogeneity, instrument relevance, and exogeneity (Bastardo et al., 2023). To assess the overall quality of the studies using an IV design, we followed recommendations by Semadeni et al. (2014), Bastardo et al. (2023), and Larcker and Rusticus (2010) and coded whether studies (a) report the results of instrument strength tests (e.g., F-test); (b) use plausibly exogenous instruments applying the classification developed by Bastardo et al. (2023); and (c) discuss the exclusion restriction, which means that the authors discuss why it is implausible to assume that the instrument is related to the dependent

variable through any other channel than the endogenous treatment. An overview of the coding can be found in Appendix B.

Regarding the reported information, ten of 16 studies (62.5 percent) provide information regarding instrument strength (e.g., OLS-F-statistic), which means that more than one-third of the studies omit this crucial information. Of those ten studies that report instrument strength, the average F-value was 43.65 (maximum 119.88; minimum 10.51), which suggests a sufficiently high instrument strength.

Only four studies (25.0 percent) report the results of an endogeneity test (e.g., Wu-Hausman test), which compares the results of an efficient but probably inconsistent estimator (OLS) with the results of an inefficient but consistent estimator (IV; only under the assumption of a valid instrument) (Bastardo et al., 2023). A significant test shows that OLS and IV estimates differ, which indicates the (probable) inconsistency of the OLS estimator and suggests that the inefficient but consistent IV estimator should be used.

Regarding the applied instruments, our analysis shows that six studies (37.5 percent) use internal instruments, i.e., an instrument based on factors internal to an organization, which is likely to be endogenous and, thus, inappropriate (Bastardo et al., 2023). In contrast, nine studies (56.3 percent) use an instrument based on factors external to an organization (i.e., external instrument), which is more likely to be exogenous and, thus, more likely to be appropriate (Bastardo et al., 2023).⁷ Following Bastardo et al. (2023), we classified the degree of appropriateness of the applied instrument from “inappropriate” to “probably appropriate.” We classified an instrument as “inappropriate” if it was either endogenous or (clearly) violated the exclusion restriction.

⁷ In one study, we could not find any information regarding the applied instrument.

Table 1
Overview of Identification Strategies and Key Assumptions.

Identification Strategy	Explanation	Key Assumptions	References
Statistical Adjustment	Identifying the causal effect of a non-random treatment by statistically “controlling” for all relevant covariates	Unconfoundedness	Antonakis et al. (2010), Wooldridge (2009)
Instrumental Variable Design	Identifying the causal effect of a non-random treatment by using an exogenous instrument	Instrument relevance; instrument exogeneity; exclusion restriction	Bastardo et al. (2023), Lal et al. (2023)
Difference-in-Differences Design	Identifying the causal effect of a non-random treatment by comparing the differences in the changes in an outcome between a treatment and control group before and after the treatment	Parallel trend assumption; no-anticipation assumption	Roth et al. (2023), Wing et al. (2018)
Matching	Identifying the causal effect of a non-random treatment by matching similar subjects that received a/no treatment	(Weak) unconfoundedness;	Narita et al. (2023), Iacus et al. (2019), Sekhon (2009)
System GMM	Identifying the causal effect of a non-random treatment by using internal instruments	Error term is serially uncorrelated; instrument exogeneity	Li et al. (2021), Roodman (2009a), Ullah et al. (2018)

Table 2
Overview of Applied Identification Strategies.

Identification Strategy	Total	in percent
Statistical Adjustment	60	93.8 %
Instrumental Variable Design	16	25.0 %
difference and system GMM	11	17.2 %
Matching	2	3.1 %
Difference-in-Differences Design	1	1.6 %
Heckman Selection	0	0.0 %
Regression Discontinuity Design	0	0.0 %

N = 64 articles; articles can use multiple identification strategies.

Eleven of the 16 studies (68.8 percent) use at least one instrument that is classified as being “potentially appropriate” (e.g., shift-share instrument, pre-quota diversity ratio, gender quota law), whereas seven studies (43.8 percent) use at least one instrument that is classified as being “inappropriate” (e.g., board size, lagged board gender diversity).⁸

Finally, we analyzed whether studies discuss the exclusion restriction, a crucial assumption of the IV design (Bastardo et al., 2023). We found only eight articles (50.0 percent) in which the exclusion restriction was discussed.

To summarize, of the 16 studies using an IV design, seven (43.8 percent) report all necessary information and use an appropriate instrument so that we can classify these studies as plausibly providing causal estimates. Four articles (25.0 percent) provide insufficient information to check the plausibility of the causal effect, whereas five articles (31.3 percent) are classified as being unlikely to provide causal estimates mainly because of the use of inappropriate instruments.

Difference and system generalized method of moments (GMM)

The difference and system GMM are both estimators designed for panel data with few time periods and many cross-sections (Roodman, 2009a). Similar to the IV design, the difference and system GMM use instruments for identification. However, the instruments in the difference and system GMM are commonly internal, which means from within the dataset itself (e.g., lags of the endogenous independent variable, see Roodman, 2009a), and not external instruments, as in most IV applications.⁹ Although internal instruments are often readily available, especially in studies using panel data, they require careful testing to

⁸ Please note that a study with multiple instruments can have both “potentially appropriate” and “inappropriate” instruments.

⁹ The difference GMM addresses bias in dynamic panel data (i.e., the correlation between the lagged dependent variable and the error term) in two steps: first, by using first-differencing to remove unit-specific fixed effects. Second, by using lagged values of the dependent and independent variables as instruments for the transformed data (Roodman, 2009a). For instance, board gender diversity at time T0 is instrumented by its past values from T-1, T-2, etc. The system GMM extends this approach by additionally using first-differences of the variables as instruments, which is more efficient (Roodman, 2009a).

ensure their validity.

Our review found that eleven out of the 64 studies (17.2 percent) use difference and system GMM to estimate a causal effect between strategic leadership team diversity and firm performance. Out of these eleven studies, none use the difference GMM, whereas the remaining studies either use the system GMM (7 studies; 63.6 percent) or do not report the specific GMM estimator (4 studies; 36.4 percent).

The difference and system GMM are attractive identification strategies because they avoid one of the critical challenges of the IV design: the quest for valid (external) instruments. However, we must emphasize that the difference and system GMM are no “magic tools” that estimate causal effects under all circumstances. Roodman (2009b, p. 156) even alerted researchers that difference and system GMM may “generate results that are invalid and appear valid.” To assess the quality of the studies using difference and system GMM, we analyzed whether researchers followed recommendations in the literature (Li et al., 2021; Roodman, 2009b). Specifically, we analyzed whether they (a) report the number of instruments, (b) test the sensitivity of the results to reductions in the number of instruments, (c) test instrument validity (e.g., by using Hansen J test or Sargan test), (d) test the validity of a subset of instruments (e.g., via the difference-in-Hansen tests), and (e) test for autocorrelation (e.g., by using Arellano-Bond test). An overview of the coding can be found in Appendix C.

Overall, we identify considerable potential for improving reporting practices: Of the eleven studies using difference and system GMM, none report the number of instruments, and no study reports sensitivity tests for a reduced number of instruments. We find more encouraging results regarding instrument exogeneity, where nine out of eleven studies (81.8 percent) report the results of tests, such as the Hansen J-test or Sargan test. Yet, no study reports the results of a difference-in-Hansen/Sargan test that tests the validity of a subset of instruments. Finally, eight studies (72.7 percent) present the result of autocorrelation tests, mainly the Arellano-Bond test.

To summarize, no study using the difference and system GMM provides sufficient information for readers to draw causal conclusions, which is essential given the reliance on internal instruments, which are prone to violating the necessary assumptions. Given the potential of both difference and system GMM to provide false results (e.g., Roodman, 2009b), we conclude that the evidence regarding the relationship between strategic leadership team demographic diversity and firm performance presented in these studies must be interpreted cautiously.

Matching methods

Matching is a statistical technique used in observational studies to estimate causal effects. It aims to overcome the non-random assignment of a treatment by matching similar subjects in the treatment and control groups. Matching has seldom been applied in research on the relationship between strategic leadership team diversity and firm performance; only two articles (3.1 percent) used this approach.

To evaluate the quality of research on strategic leadership team demographic diversity using matching, we followed recommendations in

the literature (Stuart, 2010). We analyzed whether these studies (a) report balance tests between treatment and control groups that show to what extent the two groups differ, (b) discuss the unconfoundedness assumption, and (c) report sensitivity analyses. Our coding can be found in Appendix D. None of the two studies reports any of the required information. Therefore, we classified both studies as providing “insufficient information.”

The difference-in-differences design (DID design)

The DID design can be applied in situations in which treatment and control groups are observed both before and after the treatment assignment. To estimate the causal effect of the treatment, researchers analyze the difference in the change in the outcome variable from the pre-treatment period to the post-treatment period between the treatment and control group (Imbens & Wooldridge, 2009).

The DID design is one of the most frequently applied identification strategies in empirical work using observational data. However, in research on the effect of strategic leadership team demographic diversity on firm performance, only one study (Yang et al., 2019) in our sample (1.6 percent) used the DID design. Causal identification using the DID design is based on several key assumptions (Roth et al., 2023; Wing et al., 2018): First, it is assumed that trends in the outcome variable (e.g., firm performance) would have been similar for the treatment and control group in the absence of a treatment (i.e., common trend or parallel trend assumption). This assumption is violated if treatment and control groups differ regarding unobserved time-varying factors correlated with pre-treatment trends in the outcome (e.g., companies in treatment and control groups differ regarding their organizational culture, which changes over time and affects firm performance). Second, we assume no anticipation of the treatment effects. This means that subjects (e.g., firms) do not change their behavior before the treatment is implemented (e.g., companies in the treatment group anticipate being affected by a board gender quota and, thus, decide to change board composition even before the quota is in effect). We coded how the study dealt with these assumptions using the DID design. An overview of our coding can be found in Appendix E.

Regarding the parallel trend assumption, the article discusses the issue and a figure showing the trends in firm performance over time for companies in the treatment and control groups. Regarding strict exogeneity, the authors discuss the potential for anticipation effects and selective responses to the quota. To conclude, given that the article checks the plausibility of the key assumptions of the DID design, we classify the article as reporting “plausible causal effects.”

Part II: The causal effect of strategic leadership team demographic diversity on firm performance

Overall, our review supports the recent calls by several strategic leadership scholars (e.g., Neely et al., 2020; Vera et al., 2022) for conclusive causal evidence. In this section, we review the current causal evidence regarding the relationship between strategic leadership team demographic diversity and firm performance. We classified only seven articles (10.9 percent) as providing “plausible causal evidence.” We discuss these articles in detail in this section. Please note that we excluded all other studies in this discussion, including those that provide insufficient information for a thorough assessment, because of the uncertainty regarding the quality of the causal evidence. We structure the section around the evidence for (a) the TMT and (b) the board of directors. Table 3 provides an overview of the studies and their key results. Following Hamann and colleagues (2013), we classified firm performance variables according to different dimensions of organizational performance.

The causal effect of TMT demographic diversity on firm performance

Based on our literature review on the relationship between top management team (TMT) demographic diversity and firm performance,

we identified one study (Sieweke et al., 2023) that provided plausible causal evidence. The authors use data from S&P 1,500 firms over a 24-year timespan to analyze the effect of TMT gender diversity on firm performance. To address the non-random assignment of women to TMTs, the authors use an IV design in which they use a so-called shift-share instrument. This instrument combines the (exogenous) growth in TMT gender diversity among all firms within an industry (*shift part*) with a pre-determined firm-specific TMT gender diversity ratio (*share part*). Although the instrument does not meet the highest quality standards because it lacks experimental randomization (Bastardo et al., 2023), it relies on weaker assumptions for causal identification—meaning fewer or less stringent conditions need to hold true for causal identification—than commonly used instruments, such as the industry average of the endogenous treatment (e.g., industry average TMT gender diversity). Overall, the authors find that TMT gender diversity positively affects firms’ profitability, liquidity, and growth but has no impact on market-based performance. Also, their moderator analysis provides no evidence that the effect of TMT gender diversity on firm performance is stronger during an economic crisis. Overall, we argue that this study – despite its limitations (e.g., a potential violation of the exclusion restriction, which is indicated by the larger size of the IV estimates as compared to the OLS estimates) – provides initial evidence for a positive causal effect of TMT gender diversity on firm performance.

The causal effect of board demographic diversity on firm performance

We identified six articles that provide plausible causal evidence regarding the influence of board demographic diversity on firm performance. Most of these articles (n = 4) focus on the effect of board gender diversity. These articles exploit exogenous shocks, such as the Italian board gender quota (Ferrari et al., 2022) and the Norwegian gender quota (Yang et al., 2019), to estimate the causal effect of board gender diversity on firm performance.¹⁰ All articles assume omitted variable bias to be the main source of endogeneity and they address this bias by applying IV (Adams & Ferreira, 2009; Ferrari et al., 2022; Havrylyshyn et al., 2023) and difference-in-differences designs (Yang et al., 2019).

Overall, these four articles point towards a non-significant effect of board gender diversity on firm performance. For instance, Ferrari and colleagues (2022) find non-significant effects for all seven of their firm performance measures. This finding is supported by the studies of Adams and Ferreira (2009) and Havrylyshyn and colleagues (2023). Yang and colleagues (2019) provide mixed evidence, with their difference-in-differences study finding no significant effect of board gender diversity on the market-to-book ratio and Tobin’s Q but negative effects on operating income and ROA.

Whereas the evidence for board gender diversity points towards no effect on performance, the evidence regarding ancestral diversity and genetic diversity is more conclusive, although we must consider that both effects have been researched only once. Regarding ancestral diversity, Giannetti and Zhao’s (2019) IV design finds a positive and significant impact on two measures of firm performance (earnings per share and Tobin’s Q).¹¹ Delis and colleagues (2017) analyze the effect of board genetic diversity on firm performance. The authors are concerned with three potential sources of endogeneity: omitted variable bias,

¹⁰ As highlighted by one of our reviewers, quotas can be regarded as a contextual boundary condition for the influence of strategic leadership team demographic diversity on firm performance because we may assume that members of strategic leadership teams may react differently to the presence of a group member who is in the team because of a quota (Morgenroth & Ryan, 2018). We believe this is an interesting perspective that deserves further research.

¹¹ The authors do not explain in detail which potential source of endogeneity the IV design addresses. They even argue that they initially expected ancestral diversity to be not endogenous (Giannetti & Zhao, 2019, p. 1134).

Table 3
Overview of Causal Effects of Strategic Leadership Team Demographic Diversity on Firm Performance.

TMT						
Diversity Category	Dependent Variable	Result (Direct effect)	Moderator	Result (Moderation)	Mechanism	Reference
Gender	Profitability (return on assets)	positive			not tested	Sieweke et al. (2023)
Gender	Stock market performance (total shareholder return)	non-significant			not tested	Sieweke et al. (2023)
Gender	Liquidity (cash flow return on assets)	positive			not tested	Sieweke et al. (2023)
Gender	Growth (sales)	positive			not tested	Sieweke et al. (2023)
Gender	Profitability (return on assets)		Crisis	non-significant	not tested	Sieweke et al. (2023)
Gender	Market-based performance (total shareholder return)		Crisis	non-significant	not tested	Sieweke et al. (2023)
Gender	Liquidity (cash flow return on assets)		Crisis	non-significant	not tested	Sieweke et al. (2023)
Gender	Growth (sales)		Crisis	non-significant	not tested	Sieweke et al. (2023)
Board of Directors						
Diversity Category	Dependent Variable	Result (Direct Effect)	Moderator	Result (Moderation)	Mechanism	Reference
Gender	Stock market performance (Tobin's Q)	non-significant			not tested	Adams & Ferreira (2009)
Gender	Profitability (return on assets)	non-significant			not tested	Adams & Ferreira (2009)
Gender	Profitability (operating income/assets)	negative			not tested	Yang et al. (2019)
Gender	Profitability (return on assets)	negative			not tested	Yang et al. (2019)
Gender	Stock market performance (market-to-book ratio)	non-significant			not tested	Yang et al. (2019)
Gender	Stock market performance (Tobin's Q)	non-significant			not tested	Yang et al. (2019)
Gender	Growth (number of employees)	non-significant			not tested	Ferrari et al (2022)
Gender	Profitability (return on assets)	non-significant			not tested	Ferrari et al (2022)
Gender	Profitability (profits)	non-significant			not tested	Ferrari et al (2022)
Gender	Stock market performance (Tobin's Q)	non-significant			not tested	Ferrari et al (2022)
Gender	Operational performance (production)	non-significant			not tested	Ferrari et al (2022)
Gender	Growth (assets)	non-significant			not tested	Ferrari et al (2022)
Gender	Liquidity (short-term debts)	non-significant			not tested	Ferrari et al (2022)
Gender	Profitability (return on assets)	non-significant			not tested	Havrylyshyn et al. (2023)
Gender	Profitability (net income)	non-significant			not tested	Havrylyshyn et al. (2023)
Gender	Profitability (return on assets)		Formal board gender contact intensity	positive	not tested	Havrylyshyn et al. (2023)
Gender	Profitability (net income)		Formal board gender contact intensity	positive	not tested	Havrylyshyn et al. (2023)
Gender	Profitability (return on assets)		men directors' critical mass	non-significant	not tested	Havrylyshyn et al. (2023)
Gender	Profitability (net income)		gender contact history	non-significant	not tested	Havrylyshyn et al. (2023)
Gender	Profitability (net income)		men directors' critical mass	non-significant	not tested	Havrylyshyn et al. (2023)
Gender	Profitability (net income)		gender contact history	non-significant	not tested	Havrylyshyn et al. (2023)
Ancestral diversity	Profitability (earnings per share)	positive			Corporate risk taking	Giannetti & Zhao (2019)
Ancestral diversity	Stock market performance (Tobin's Q)	positive			Corporate risk taking	Giannetti & Zhao (2019))
Genetic diversity	Stock market performance (Tobin's Q)	positive				Delis et al. (2017)
Genetic diversity	Profitability (risk-adjusted returns)	positive				Delis et al. (2017)
Genetic diversity	Growth (sales)	positive				Delis et al. (2017)
Genetic diversity	Operational performance (operating expenses)	non-significant				Delis et al. (2017)

Notes: The overview only includes studies providing plausible causal effects of the relationship between strategic leadership team demographic diversity and firm performance.

reverse causality, and measurement error. Their IV design addresses all three sources and provides evidence for a positive effect of board genetic diversity on Tobin's Q, risk-adjusted returns, and sales growth but a non-significant effect on operating expenses.

Overall, the articles' findings indicate that different types of demographic diversity can affect firm performance differently. Whereas the evidence for board gender diversity suggests a non-significant effect, we have more conclusive evidence for a positive effect of ancestral diversity and genetic diversity. However, we must consider that this evidence is currently based on the result of a single study.

Recommendations for future research

In the following section, we will set the agenda for future research on strategic leadership team demographic diversity organized under three headings: (1) improving causal identification, (2) causally testing mechanisms and boundary conditions, and (3) broadening the focus on demographic diversity. The first recommendation reflects on the question of whether demographic diversity in strategic leadership teams affects firm performance, whereas the second recommendation reflects on the question of why and when we can observe an effect. Finally, the third recommendation addresses the focus on gender diversity in current research on strategic leadership team demographic diversity.

Improving causal identification

Our review reveals that only a few studies offer plausible causal evidence on the relationship between strategic leadership team demographic diversity and firm performance. However, causal evidence is essential for informing policy-makers and testing theory (Antonakis, 2017). This lack of plausible causal evidence is because researchers (a) apply unsophisticated identification strategies, i.e., identification strategies that rely on strong assumptions, and (b) apply sophisticated identification strategies without sufficient care, i.e., providing insufficient information to evaluate the identification strategy. Specifically, only 24 of the 64 articles in our sample (37.5 percent) use at least one sophisticated identification strategy. Further, many studies either need to report more information to assess whether the identification strategies provide plausible causal estimates or violate assumptions of the identification strategy in the application (e.g., use of an inappropriate instrument). This section provides hands-on recommendations for improving causal identification, emphasizing the need to (a) use identification strategies better and (b) use better identification strategies.

Use identification strategies better

Using identification strategies better means that researchers should carefully explore whether key assumptions for the applied identification strategies are met. Many studies lack sufficient information on the plausibility of the assumptions, hindering the assessment of the quality of the causal evidence. We offer suggestions for three common strategies: the IV design, the difference and system GMM, and matching methods. Following these recommendations will enhance transparency and enable reviewers and readers to evaluate the quality of causal evidence better.

Instrument variable design (IV Design). A crucial point in applying the IV design is to test the plausibility of the design's identifying assumptions. Yet, our review shows that few studies report the results of statistical tests required to evaluate the quality of the IV design. To improve future research, we recommend the following:

1. **Select instrument(s) carefully.** Identifying relevant instruments is complex and requires deep knowledge of the study's institutional context (Angrist & Pischke, 2009)—for instance, Ferrari et al. (2022) use the staggered introduction of a board gender quota as an

instrument for gender diversity. This instrument is exogenous and strongly affects the endogenous treatment variable (i.e., board gender diversity), which makes it a valid instrument. Whereas laws and quotas are often valid instruments, instruments such as lagged values of the endogenous variable or industry averages (e.g., the average industry-level board diversity) are likely invalid because they combine the endogenous and exogenous parts of the original variables (Larcker & Rusticus, 2010).

2. **Provide information regarding instrument strength.** Many studies lack information regarding instrument strength. Given that IV estimates are biased and inconsistent in the presence of weak instruments (Bastardo et al., 2023), this information is essential to assess the quality of the causal evidence. Therefore, we urge research to report instrument strengths. At the same time, researchers must ensure that the reported test for instrument strength fits data characteristics. For instance, the F-value is inapplicable in non-homoscedastic settings and should be replaced by the Montiel Olea and Pflueger (2013) F-statistic, which is robust to autocorrelation, heteroscedasticity and clustering (Andrews et al., 2019).¹² Given that most studies in our sample use panel data, which violate the independent and identically distributed (iid) data assumption, we recommend using the Montiel Olea and Pflueger (2013) F-statistic.
3. **Compare IV with OLS estimates.** Although not a formal test of an assumption of the IV design, we recommend that researchers compare IV and OLS estimates. The idea behind using the IV design in research for the relationship between strategic leadership team demographic diversity and firm performance is often to avoid an upward bias in the OLS estimates (i.e., to exaggerate the effect of strategic leadership team demographic diversity). Yet, some studies show IV estimates to be much larger (> three times) than OLS estimates (e.g., Sieweke et al., 2023). As Lal et al. (2023) discuss, this discrepancy can be due to heterogeneous treatment effects and violation of the exogeneity assumption. We recommend researchers check the plausibility of the IV estimates compared to the OLS estimates by comparing their size.
4. **Take the exclusion restriction seriously.** The exclusion restriction is often neglected in research on strategic leadership team demographic diversity, although IV estimates are severely biased if it is violated (Bastardo et al., 2023). Because the exclusion restriction is statistically untestable (Morgan & Winship, 2015), we recommend (a) explaining why the instrument affects the outcome only through the endogenous treatment and (b) using sensitivity tests that explore the robustness of the results concerning potential violations of the exclusion restriction (Conley et al., 2012).
5. **Further readings.** We recommend researchers consult the introduction to the IV design by Bastardo and colleagues (2023) and Lal and colleagues' (2023) review of IV applications in political science as further readings.

Difference and system GMM. The difference and system GMM use internally generated instruments to address endogeneity, making them suitable for analyzing the effect of strategic leadership team demographic diversity on firm performance if panel data with multiple periods are available. However, they have limitations, and there is a risk that results that appear to be valid are, in fact, invalid (Roodman, 2009b). To mitigate this risk, we recommend reporting the following information (e.g., Li et al., 2021; Roodman, 2009a, 2009b):

1. **Check for autocorrelation.** Researchers should test whether the error terms are serially uncorrelated, e.g., using the Arellano-Bond test, because serial correlation would render the instruments' lags invalid (Roodman, 2009a). Whereas we can expect to reject the null

¹² Please note that the commonly used Stock and Yogo (2005) critical F-values should not be used in non-homoscedastic settings.

- hypothesis at lag 1 of the instruments, lag 2 should be insignificant (ideally with a high p-value) (Li et al., 2021).
- Report the number of instruments.** Given that a large number of instruments compared to the number of observations can bias estimates (Roodman, 2009b), we recommend researchers report the number of instruments. Also, researchers should test whether their results are robust to a reduced number of instruments (e.g., by using a lower number of lags or by collapsing instruments; see Roodman, 2009b).
 - Check instrument exogeneity.** Researchers should check instrument exogeneity, e.g., by using the Hansen J-test. The Hansen J-test is sensitive to the number of instruments, which is vital given that a large number of instruments can lead to implausible p-values of 1.00 (Roodman, 2009b). Therefore, researchers should carefully check the Hansen J-test of their full model against the result of a J-test with a reduced set. In the case of the system GMM, researchers should also test the exogeneity of the additional subsets of the instruments (e.g., the transformed equation and levels equation), e.g. using the difference-in-Hansen/Sargan test (Li et al., 2021).
 - Further readings.** We recommend that researchers familiarize themselves with more specialized literature (e.g., Li et al., 2021; Roodman, 2009a, 2009b) before applying the difference and system GMM.

Matching methods. Matching methods rely on strong assumptions, most notably the unconfoundedness assumption (Imbens, 2015). Researchers should analyze the plausibility of these assumptions when applying matching methods. We recommend the following analyses when using matching methods (e.g., Imai et al., 2023; Imbens, 2015; Narita et al., 2023):

- Test the plausibility of the unconfoundedness assumption.** Although the unconfoundedness assumption is not testable, researchers can conduct tests to provide supportive evidence. After matching, Imbens (2015) recommends estimating the effect of the treatment on a pseudo-outcome, i.e., an outcome where researchers know that it is unaffected by the treatment (e.g., a pre-treatment measure of the outcome). Suppose the treatment is related to the pseudo-outcome. In that case, the unconfoundedness assumption is probably violated, and matching methods should not be applied, whereas an effect close to zero indicates that the assumption is more plausible.
- Check covariate balance.** Researchers should check the overlap in covariate distribution between treatment and control groups, e.g., by calculating normalized differences in covariates between treatment and control groups (Imbens & Wooldridge, 2009). Imbens (2015) suggests dropping units with no counterparts in the other group if the treatment and control groups differ regarding the normalized difference metric. He also discusses a rule for dropping extreme units, which is based on calculating a propensity score and, thus, gives less room for researcher subjectivity.
- Test the sensitivity of the results.** Although researchers can never rule out the possibility of unmeasured confounders, they can analyze the behavior of the treatment effect for different specifications of the unobserved confounders (Steiner & Cook, 2013). For instance, they can test what happens if the unmeasured confounder has the same property as important observed covariates or how strongly an unmeasured confounder must be correlated with the treatment variable and the outcome for the treatment effect to disappear. Based on the results of the sensitivity analyses, researchers can discuss whether it is plausible to assume that a confounder with such properties is omitted (Ichino et al., 2008). Also, researchers should examine how changes in the specification of the propensity score affect results (Dehejia, 2005). This analysis provides further insights into the plausibility of a causal interpretation of the estimates.

- Further readings.** We suggest interested researchers consult more specialized literature before applying matching methods (Imai et al., 2023; e.g., Imbens, 2015; Narita et al., 2023; Stuart, 2010).

Use better identification strategies

Besides using identification strategies better, research on strategic leadership team demographic diversity can be improved by using better identification strategies. Our review shows that many studies use identification strategies (e.g., statistical adjustment) with strong assumptions that are often violated. We recommend using strategies that require weaker assumptions for causal identification.

To support researchers in selecting “better” identification strategies, we developed a decision tree (see Fig. 2) structured around critical questions regarding the demographic diversity variable. The first question is whether diversity is exogenously manipulated, e.g., through quotas. If “yes,” then researchers need to analyze whether the treatment assignment is based on an observed variable, e.g., a company’s number of employees. If the answer is “yes,” researchers may apply an RDD. If the answer is “no,” then they need to analyze whether companies must comply with the treatment, e.g., in the case of a “hard” quota, or whether companies can decide whether they comply with the treatment, e.g., in the case of a “soft” quota. If companies must comply, researchers should check whether they find a comparable control group. For instance, board gender quotas often only apply to public companies. The unaffected companies (e.g., private companies) can be used as a control group, and researchers can use the DID design (e.g., Yang et al., 2019). If companies do not need to comply with the treatment or there is no comparable control group, we recommend using the IV design with the quota as an instrument (e.g., Ferrari et al., 2022).

If diversity is not exogenously manipulated, researchers must analyze whether a valid external instrument is available (e.g., ancestral diversity within a country, see Giannetti & Zhao, 2019). If available, they must decide whether endogeneity results from unobserved selection processes, that is, some values of the dependent variable are missing because of a non-random selection of observations into the sample—then use Heckman selection. If endogeneity results from other sources, such as omitted variable bias—then an IV design must be used because the Heckman selection models produce biased results in the case of omitted variables (Certo et al., 2016).¹³ If no valid external instruments are available, researchers may consider using internal instruments, such as when using the difference or system GMM. Finally, if internal instruments are invalid, researchers may rely on matching methods requiring a rich set of covariates (Smith & Todd, 2005).

We hope this decision tree aids in selecting appropriate identification strategies to analyze the causal relationship between strategic leadership team demographic diversity and firm performance. Although we included many identification strategies in the decision tree, we prefer the RDD and the DID design because they rely on relatively weak assumptions for causal identification. To facilitate their use in empirical research, we describe both in an accessible way with examples.

Difference-in-Differences design (DID design). Despite its prominence in fields including economics (Currie et al., 2020) and political science (Hassell & Holbein, 2024), the DID design has seldom been used in research on the relationship between strategic leadership team demographic diversity and firm performance (for an exception, see Yang et al., 2019). We encourage using the DID design because it is well-suited to estimate causal effects. For instance, several countries have discussed

¹³ The inability of the Heckman model to resolve endogeneity from other sources than sample selection bias makes us believe that this identification strategy should be used only very carefully in research on strategic leadership team demographic diversity. We rather recommend using the IV design, which addresses endogeneity resulting from omitted variable bias (Bastardo et al., 2023).

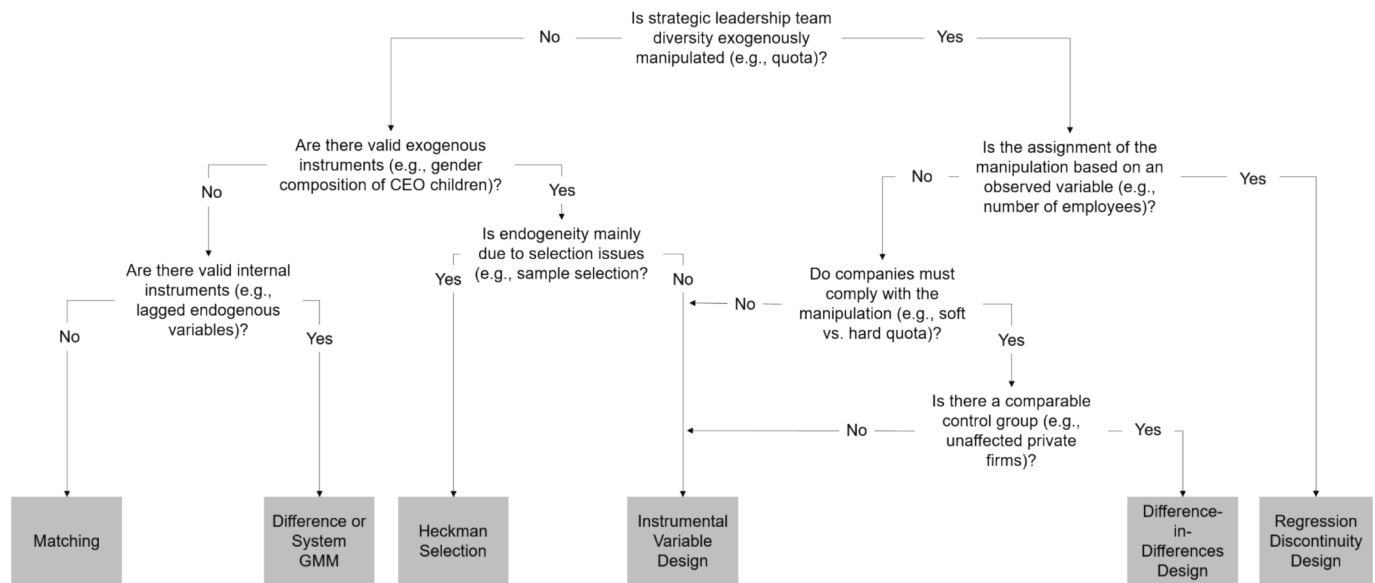


Fig. 2. Decision Tree for Selecting an Identification Strategy.

(e.g., the Netherlands and Germany) or have introduced gender quotas in TMTs and corporate boards (e.g., Norway). Also, California has recently introduced bills to increase the representation of women (Senate Bill No. 826) and underrepresented communities (Assembly Bill No. 979), such as African Americans and Latinos, in corporate boards. The DID design can be used in these contexts to compare affected firms (i.e., treatment group) with unaffected firms (i.e., control group), such as firms in other countries or states (e.g., Matsa & Miller, 2013; Yang et al., 2019).

To ensure that future studies exploit the potential of the DID design, we recommend the following (e.g., Cunningham, 2021; Roth et al., 2023; Wing et al., 2018):

- Carefully select a control group.** The first important step is choosing a control group that closely matches the treatment group. Whereas identifying the treatment group is often straightforward, finding a comparable control group can be challenging. For instance, companies affected by the Californian board gender quota (State of California, 2018) represent the treatment group, yet who is the control group? Researchers have different options: they may select private Californian companies because only public companies are affected by the bill.¹⁴ Alternatively, they can compare public companies in California with public companies in U.S. states without a quota. Alternatively, researchers may apply a “difference-in-differences” design (Olden & Møen, 2022), comparing performance differences between public and private companies in California with differences between public and private companies in other U.S. states (for a similar approach, see Matsa & Miller, 2013). Regardless of the choice, researchers must ensure that the treatment and control groups are as similar as possible before the treatment to prevent confounding effects and justify their choice.
- Test the parallel trend assumption.** The parallel trend assumption, which indicates that the trend in the outcome for the treatment and control groups “would have evolved in parallel if treatment had not occurred,” (Roth et al., 2023, p. 2221) is a key identifying assumption of the DID design. For instance, differences in pre-treatment trends in firm performance between public (treatment group) and private companies with headquarters in California (control group)

may indicate the presence of unobserved time-varying characteristics (e.g., company culture). We must emphasize that the parallel trend assumption allows for non-random treatment assignment (e.g., self-selection effects) if these characteristics only affect the outcome level but not the trend (Roth et al., 2023). Testing the parallel trend assumption requires data from at least two pre-treatment periods; yet, we recommend collecting data from as many pre-treatment periods as possible to increase statistical power to detect a potential violation. We suggest two steps to test the parallel trend assumption: first, researchers should visually inspect the pre-treatment trends to identify a non-parallel movement (Yang et al., 2019). Second, we recommend applying statistical tests to identify potential violations. The most common statistical test regresses the outcome variable on interaction terms of the treatment variable and the pre-treatment periods (dummy coded) (Kahn-Lang & Lang, 2020). However, these tests often lack statistical power to detect parallel trend violations (Roth, 2022). Roth (2022) proposes potential solutions to the problem. For now, we recommend researchers test both the individual and joint significance of the estimates to detect possible violations. The DID design can still be applied even if the parallel trends are violated. Rambachan and Roth (2023) developed a tool for robust inference, and O’Neill and colleagues’ (2016) simulations show that the lagged dependent variable regression approach offers the least biased estimates under such conditions.

- Expect anticipation effects.** Researchers should understand the research context to detect potential anticipatory behavior. Two types are possible: First, companies may already change the diversity of their board in anticipation of the law (i.e., in the pre-treatment period), influencing performance before the treatment and, thereby, violating the strict exogeneity assumption (Wing et al., 2018). Second, anticipatory behavior may change the composition of the treatment and control group from the pre-treatment to the post-treatment period. For instance, public companies may move from California to Texas to avoid being affected by the new bill. To address this problem, researchers should use the law’s announcement to define pre- and post-treatment periods. Therefore, we recommend robustness checks, e.g., analyzing effects if the announcement of the law is defined as a post-treatment period, and additional analyses, e.g., checking changes in the composition of the treatment and control group over time, to identify potential anticipatory behaviors.
- Further readings.** To better understand the DID design, researchers should consult specialized literature (e.g., Roth et al., 2023; Wing

¹⁴ This approach would follow studies on the effects of the Norwegian gender quota in corporate boards (e.g., Matsa & Miller, 2013; Yang et al., 2019).

et al., 2018). Recent advancements in the DID-related econometrics literature, such as staggered treatment assignment (Goodman-Bacon, 2021), and future developments, e.g., regarding DID with continuous treatments (Callaway et al., 2024), are expected to be relevant for research on strategic leadership team demographic diversity.

Regression discontinuity design (RDD). The RDD is the quasi-experimental design with the weakest identifying assumptions because treatment assignment depends on subjects' scores on an observed variable (Lee & Lemieux, 2010). In fact, the RDD may even resemble the internal validity of the randomized experiment (Antonakis et al., 2010). Therefore, we recommend researchers apply the RDD whenever possible to estimate the causal effect of strategic leadership team demographic diversity on firm performance with observational data.

Although the RDD is a powerful identification strategy, it has yet to be used in research on strategic leadership team demographic diversity. Applying the RDD requires that companies receive a treatment based on a score on an observed variable. For instance, policy-makers may introduce a law that forces all companies with more than 250 employees to have at least one woman on their board or TMT, whereas companies below this cut-off are unaffected. Researchers can exploit this setup to estimate the effect of gender diversity on firm performance by comparing companies just above and below the cut-off of 250 employees. The key identifying assumption of the RDD is that close to the cut-off, treatment assignment is "as-if random." That is, companies with 249 and 250 employees do not systematically differ from each other except for the treatment.

Regarding empirical settings where the RDD can be applied, we believe the Swiss gender quota in corporate boards provides exciting opportunities. The Swiss government requires "large" companies to achieve a gender quota of 30 percent for corporate boards and 20 percent for TMTs (Der Bundesrat: Das Portal der Schweizer Regierung, 2020).¹⁵ An RDD is potentially applicable because "large" companies are defined based on whether companies score above specific criteria regarding their balance sheet, sales revenue, and number of employees.

The validity of the RDD relies on certain assumptions, so researchers must carefully assess their plausibility. We recommend conducting at least three analyses, as suggested in prior research (e.g., Bastardo et al., 2024; Cattaneo & Titiunik, 2022; Sieweke & Santoni, 2020):

- 1. Test the plausibility of as-if randomization.** Researchers should analyze the plausibility of an "as-if random" treatment assignment around the cut-off. If this assumption holds, companies in treatment and control groups near the cut-off should not systematically differ (Cattaneo & Titiunik, 2022). To evaluate the plausibility of the assumption, researchers can use balance tests, such as normalized differences (e.g., Imbens & Wooldridge, 2009).
- 2. Critically evaluate potential manipulations around the cut-off.** A fundamental assumption of the RDD is that subjects cannot precisely manipulate the treatment assignment (Lee & Lemieux, 2010). Analyzing potential manipulation around the cut-off is crucial for internal validity because active manipulation could cause differences between those subjects who actively manipulate and those who do not. To detect potential manipulations, researchers should use McCrary's (2008) test, which analyzes whether the density of the assignment variable is continuous or discontinuous around the cut-off or Cattaneo et al.'s (2020) density estimator. A discontinuous density function (i.e., a "jump") around the cut-off indicates potential manipulations, undermining the validity of the RDD.
- 3. Try to falsify the model.** Researchers should try to disprove the validity of their RDD by conducting placebo tests, i.e., tests where researchers would expect to find no effect (e.g., Cattaneo & Titiunik,

2022; Sieweke & Santoni, 2020). For instance, testing for treatment effects at placebo cut-offs, where no treatment effect should be present, can reveal potential confounding effects. Cattaneo and colleagues (2022) suggest using cut-off points above or below the actual cut-off.

- 4. Further readings.** These recommendations cover only the most essential aspects when conducting an RDD. We recommend that interested researchers consult the more specialized literature on RDD (e.g., Cattaneo et al., 2019; Cattaneo & Titiunik, 2022; Lee & Lemieux, 2010).

Causally testing mechanisms and boundary conditions

Throughout our review, we argued that it is an important first step for researchers to provide evidence for the causal relationships between strategic leadership team diversity and firm performance. As an important next step, when causality can be established, we recommend that researchers propose and rigorously test the potential theoretical mechanisms and boundary conditions that might explain these causal relationships. For policy-makers and researchers alike, it is not only important to gain a better understanding of *whether* diverse strategic leadership teams causally affect firm performance but also *why* and *when* we can observe this effect. Our systematic review reveals that prior research utilizes a wide array of theoretical perspectives and models from different disciplines such as organizational behavior, corporate governance, or social psychology. These perspectives and models include the resource perspective or the categorization-elaboration model, and theories such as the upper echelons theory, contact theory, critical mass/tokenism theory or gender role theory. These frameworks are often employed to argue, for example, for a positive effect of leadership team diversity on firm outcomes. However, the large number of theoretical approaches suggests a need for a comprehensive theoretical framework capable of explaining the inconsistent results regarding the relationship between strategic leadership team demographic diversity and firm performance. Furthermore, although most studies draw from different theories to propose mechanisms and boundary conditions of the relationship between board diversity and firm performance, they are often not explicitly and empirically tested. For example, Havrylyshyn et al. (2023) integrate core assumptions of contact theory and critical mass theory and empirically test the moderating effect of formal contact intensity among men and women director colleagues on the relationship between gender-diverse boards and firm performance. Their results demonstrate that contact between men and women directors enhances the proposed relationship between board gender diversity and firm performance. They further find that firm performance decreases when a gender-diverse board is comprised of more men who have exclusively worked on boards with token numbers of women. However, other studies (e.g., Ferrari et al., 2022; Yang et al., 2019) that do not explicitly propose theoretical mechanisms but test the effects of gender quotas for boards reveal ambiguous results on whether female representation has a positive or negative effect on firm performance.

Regarding boundary conditions, Sieweke et al. (2023) tested the proposition of upper echelons theory that the effect of TMT gender diversity on firm performance is more positive in turbulent environments resulting from an exogenous shock (Jacquart et al., 2024) by assessing the moderating effect of the 2008/2009 financial crises and the 2020 COVID-19 pandemic on the relationship between TMT gender diversity and firm performance. Their analyses provided no evidence for an advantage of higher levels of TMT gender diversity during a crisis when considering firm financial performance as an outcome variable. However, the authors stressed that "the absence of evidence is not evidence of absence" (p. 12) and recommended further research.

To summarize, we recommend that researchers engage in a joint effort to not only provide evidence for causality between strategic leadership team diversity and firm performance but also to propose and rigorously test a theoretical framework that may help to answer the

¹⁵ Given that the quota is currently "soft," i.e., companies do not have to comply with the quota, researchers may also consider using an IV design.

questions why and when board diversity leads to higher firm performance. Yet, it is essential to emphasize that testing mechanisms and boundary conditions require the same rigor as testing direct causal relationships. Recent work on heterogeneous treatment effects (e.g., Wager & Athey, 2018) in the field of economics can be relevant for identifying boundary conditions. Regarding causal mediation, we refer readers to Celli's (2022) review of causal mediation using quasi-experimental approaches (e.g., difference-in-difference, IV, and synthetic control designs) in economic studies and to the discussion in Wulff et al. (2023).

Broadening the focus on demographic diversity

A striking finding of our review is the dominance of research on the causal effect of strategic leadership team gender diversity – especially board gender diversity – on firm performance. Indeed, of the seven studies we classified as providing plausible causal evidence, four focused on board gender diversity, one focused on TMT gender diversity, and only two studies (Delis et al., 2017; Giannetti & Zhao, 2019) focused on non-gender related demographic diversity types (i.e., ancestral diversity and genetic diversity). Of course, there are good reasons for the predominance of gender-related research, mainly the implementation of gender quotas for boards and TMTs in many countries (e.g., Norway and the Netherlands). However, we can observe that policy-makers have started broadening their diversity perspective. For instance, in 2018, California passed legislation requiring publicly traded companies to have a minimum number of female directors on their boards (State of California, 2018). Then, in 2020, legislators of California extended this requirement to include a minimum number of directors from under-represented groups, such as African Americans, Latinos, or individuals who identify as gay, lesbian, bisexual, or transgender (State of California, 2020). Since policy-makers use the “business case” argument again to justify such quotas, we deem it important to investigate whether there is evidence for a causal effect on firm performance. These quotas also provide new opportunities for researchers to broaden their perspective and causally test the effect of a broader range of demographic diversity characteristics on firm performance.

Conclusion

Our systematic review of the methodologies employed by strategic

Appendix A. Overview of coded studies

No.	Authors	Journal	Year	Statistical adjustment	Matching	IV	GMM	RDD	DID	Event study	Heckman selection	Causal effect
1	Dwyer, S; Richard, OC; Chadwick, K	Journal Of Business Research	(2003)	Yes	No	No	No	No	No	No	No	Implausible causal effect
2	Erhardt, NL; Werbel, JD; Shrader, CB	Corporate Governance- An International Review	(2003)	Yes	No	No	No	No	No	No	No	Implausible causal effect
3	Van der Walt, N; Ingley, C; Shergill, GS; Townsend, A	Corporate Governance- The International Journal Of Business In Society	(2006)	Yes	No	No	No	No	No	No	No	Implausible causal effect
4	Campbell, K; Mínguez-Vera, A	Journal Of Business Ethics	(2007)	Yes	No	No	No	No	No	No	No	Implausible causal effect
5	Francoeur, C; Labelle, R; Sinclair-Desgagné, B	Journal Of Business Ethics	(2007)	Yes	No	No	No	No	No	No	No	Implausible causal effect
6	McIntyre, ML; Murphy, SA; Mitchell, P	Corporate Governance- The International Journal Of Business In Society	(2007)	Yes	No	No	No	No	No	No	No	Implausible causal effect

(continued on next page)

management team demographic diversity research demonstrates that we, as a field, cannot conclusively support the business case for strategic leadership team diversity often championed in academic research and policy-making, nor can we reject it due to insufficient evidence. This lack of clarity has several implications. Firstly, policy-makers should avoid using the inconclusive business case as the primary reason for enacting diversity regulations, and instead justify regulations based on evidence of (structural or personal) discrimination, e.g., the underrepresentation of women on corporate boards (Kirsch, 2018). Secondly, the academic community must consider boards and TMTs as interdependent (i.e., the strategic leadership team) and produce more rigorous research on strategic leadership team demographic diversity, especially as policy-makers refer to these studies to inform regulations, underscoring the real-world impact of academic work. We hope our review and the recommendations we developed will contribute to more reliable causal evidence regarding the relationship between strategic leadership team gender diversity and firm performance.

CRedit authorship contribution statement

Jost Sieweke: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Tanja Hentschel:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Brooke A. Gazdag:** Writing – review & editing, Conceptualization. **Levke Henningsen:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Dutch Research Council (NWO) under project number VI.Veni.201E.060 awarded to Tanja Hentschel.

(continued)

No.	Authors	Journal	Year	Statistical adjustment	Matching	IV	GMM	RDD	DID	Event study	Heckman selection	Causal effect
7	Rose, C	Corporate Governance- An International Review	(2007)	Yes	No	No	No	No	No	No	No	Implausible causal effect
8	Adams, RB; Ferreira, D	Journal Of Financial Economics	(2009)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
9	Ahern, KR; Dittmar, AK	Quarterly Journal Of Economics	(2012)	No	No	Yes	No	No	No	No	No	Insufficient information
10	Dezso, CL; Ross, DG	Strategic Management Journal	(2012)	Yes	No	No	Yes	No	No	No	No	Implausible causal effect
11	Joecks, J; Pull, K; Vetter, K	Journal Of Business Ethics	(2012)	Yes	No	No	No	No	No	No	No	Implausible causal effect
12	Nielsen, BB; Nielsen, S	Strategic Management Journal	(2012)	Yes	No	No	No	No	No	No	No	Implausible causal effect
13	Ujunwa, A	Corporate Governance- The International Journal Of Business In Society	(2012)	Yes	No	No	No	No	No	No	No	Implausible causal effect
14	Zhang, JQ; Zhu, H; Ding, HB	Journal Of Business Ethics	(2012)	Yes	No	No	No	No	No	No	No	Implausible causal effect
15	Zhang, L	Corporate Governance- The International Journal Of Business In Society	(2012)	Yes	No	No	No	No	No	No	No	Implausible causal effect
16	Darmadi, S	Corporate Governance- The International Journal Of Business In Society	(2013)	Yes	No	No	No	No	No	No	No	Implausible causal effect
17	Isidro, H; Sobral, M	Journal Of Business Ethics	(2014)	Yes	No	No	No	No	No	No	No	Implausible causal effect
18	Arena, C; Cirillo, A; Mussolino, D; Pulcinelli, I; Saggese, S; Sarto, F	Corporate Governance- The International Journal Of Business In Society	(2015)	Yes	No	No	No	No	No	No	No	Implausible causal effect
19	Darko, J; Aribi, ZA; Uzonwanne, GC	Corporate Governance- The International Journal Of Business In Society	(2016)	Yes	No	No	No	No	No	No	No	Implausible causal effect
20	Perryman, AA; Fernando, GD; Tripathy, A	Journal Of Business Research	(2016)	Yes	No	No	No	No	No	No	No	Implausible causal effect
21	Toumi, N; Benkraiem, R; Hamrouni, A	Corporate Governance- The International Journal Of Business In Society	(2016)	Yes	No	No	No	No	No	No	No	Implausible causal effect
22	Conyon, MJ; He, LR	Journal Of Business Research	(2017)	Yes	No	Yes	No	No	No	No	No	Insufficient information
23	Delis, MD; Gaganis, C; Hasan, I; Pasiouras, F	Management Science	(2017)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
24	Chong, LL; Ong, HB; Tan, SH	Corporate Governance- The International Journal Of Business In Society	(2018)	Yes	No	No	No	No	No	No	No	Implausible causal effect
25	McGuinness, PB	Journal Of Business Ethics	(2018)	Yes	No	No	No	No	No	No	No	Implausible causal effect
26	Pucheta-Martínez, MC; Bel-Oms, I; Olcina-Sempere, G	Journal Of Business Ethics	(2018)	Yes	No	Yes	No	No	No	No	No	Insufficient information
27	Roudaki, J	Corporate Governance- The International Journal Of Business In Society	(2018)	Yes	No	No	No	No	No	No	No	Implausible causal effect
28	Giannetti, M; Zhao, MX	Journal Of Financial And Quantitative Analysis	(2019)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
29	Martínez, MDV; Rambaud, SC; Oller, IMP	Corporate Social Responsibility And Environmental Management	(2020)	No	No	Yes	No	No	No	No	No	Insufficient information
30	Triana, MD; Richard, OC; Su, WC	Research Policy	(2019)	Yes	No	No	No	No	No	No	No	Implausible causal effect

(continued on next page)

(continued)

No.	Authors	Journal	Year	Statistical adjustment	Matching	IV	GMM	RDD	DID	Event study	Heckman selection	Causal effect
31	Ullah, I; Fang, HX; Jebran, K	Corporate Governance-The International Journal Of Business In Society	(2019)	Yes	No	No	Yes	No	No	No	No	Insufficient information
32	Uribe-Bohorquez, MV; Martínez-Ferrero, J; García-Sánchez, IM	Corporate Social Responsibility And Environmental Management	(2019)	Yes	No	No	No	No	No	No	No	Implausible causal effect
33	Wang, Y; Abbasi, K; Babajide, B; Yekini, KC	Corporate Governance-The International Journal Of Business In Society	(2019)	Yes	No	Yes	No	No	No	No	No	Implausible causal effect
34	Yang, P; Riepe, J; Moser, K; Pull, K; Terjesen, S	Leadership Quarterly	(2019)	No	No	No	No	No	Yes	No	No	Plausible causal effect
35	Aldhamari, R; Nor, MNM; Boudiab, M; Mas'ud, A	Corporate Governance-The International Journal Of Business In Society	(2020)	Yes	No	No	No	No	No	No	No	Implausible causal effect
36	Fernández-Temprano, MA; Tejerina-Gaite, F	Corporate Governance-The International Journal Of Business In Society	(2020)	Yes	No	No	No	No	No	No	No	Implausible causal effect
37	Fernando, GD; Jain, SS; Tripathy, A	Journal Of Business Research	(2020)	Yes	No	Yes	No	No	No	No	No	Implausible causal effect
38	Liu, YH; Lei, LJ; Buttner, EH	Journal Of Business Research	(2020)	Yes	No	No	No	No	No	No	No	Implausible causal effect
39	Martinez-Jimenez, R; Hernández-Ortiz, MJ; Fernández, AIC	Corporate Governance-The International Journal Of Business In Society	(2020)	Yes	No	No	No	No	No	No	No	Implausible causal effect
40	Mazzotta, R; Ferraro, O	Corporate Governance-The International Journal Of Business In Society	(2020)	Yes	No	No	Yes	No	No	No	No	Insufficient information
41	Rehman, S; Orij, R; Khan, H	Corporate Social Responsibility And Environmental Management	(2020)	Yes	Yes	No	Yes	No	No	No	No	Insufficient information
42	Vairavan, A; Zhang, GP	Corporate Governance-The International Journal Of Business In Society	(2020)	Yes	No	No	No	No	No	No	No	Implausible causal effect
43	Ali, F; Wang, M; Jebran, K; Ali, ST	Corporate Governance-The International Journal Of Business In Society	(2021)	Yes	No	No	Yes	No	No	No	No	Insufficient information
44	Calabrese, GG; Manello, A	Corporate Governance-The International Journal Of Business In Society	(2021)	Yes	No	No	No	No	No	No	No	Implausible causal effect
45	Saleh, MWA; Zaid, MAA; Shurafa, R; Maigoshi, ZS; Mansour, M; Zaid, A	Corporate Governance-The International Journal Of Business In Society	(2021)	Yes	No	No	Yes	No	No	No	No	Insufficient information
46	Uyar, A; Kuzey, C; Kilic, M; Karaman, AS	Corporate Social Responsibility And Environmental Management	(2021)	Yes	No	Yes	No	No	No	No	No	Implausible causal effect
47	Veltri, S; Mazzotta, R; Rubino, FE	Corporate Social Responsibility And Environmental Management	(2021)	Yes	No	No	No	No	No	No	No	Implausible causal effect
48	Yanadori, Y; Kulik, CT; Gould, JA	Human Resource Management	(2021)	Yes	No	Yes	No	No	No	No	No	Implausible causal effect
49	Boukattaya, S; Ftiti, Z; Ben Arfa, N; Omri, A	Corporate Social Responsibility And Environmental Management	(2022)	Yes	No	No	Maybe	No	No	No	No	Implausible causal effect

(continued on next page)

(continued)

No.	Authors	Journal	Year	Statistical adjustment	Matching	IV	GMM	RDD	DID	Event study	Heckman selection	Causal effect
50	Ferrari, G; Ferraro, V; Profeta, P; Pronzato, C	Management Science	(2022)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
51	Wu, J; Richard, OC; Triana, MD; Zhang, XH	Human Resource Management	(2022)	Yes	No	Yes	No	No	No	No	No	Implausible causal effect
52	Akhter, W; Hassan, A	Corporate Social Responsibility And Environmental Management	(2023)	No	No	No	Yes	No	No	No	No	Insufficient information
53	Alodat, AY; Salleh, Z; Nobanee, H; Hashim, HA	Corporate Social Responsibility And Environmental Management	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
54	Andoh, JAN; Abugri, BA; Anarfo, EB	Corporate Governance-The International Journal Of Business In Society	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
55	Ben Fatma, H; Chouaibi, J	Corporate Governance-The International Journal Of Business In Society	(2023)	Yes	No	No	Yes	No	No	No	No	Insufficient information
56	Farooq, M; Ahmad, N	Corporate Governance-The International Journal Of Business In Society	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
57	Fayyaz, UER; Jalal, RNUD; Venditti, M; Minguez-Vera, A	Corporate Social Responsibility And Environmental Management	(2023)	Yes	No	No	Yes	No	No	No	No	Insufficient information
58	Foster, BP; Manikas, AS; Kroes, JR	Corporate Social Responsibility And Environmental Management	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
59	Gharbi, S; Othmani, H	Corporate Governance-The International Journal Of Business In Society	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
60	Havrylyshyn, A; Schepker, DJ; Nyberg, AJ	Journal Of Business Ethics	(2023)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
61	Khatri, I	Corporate Social Responsibility And Environmental Management	(2023)	Yes	No	Yes	Yes	No	No	No	No	Insufficient information
62	Sieweke, J; Bostandzic, D; Smolinski, SM	Leadership Quarterly	(2023)	Yes	No	Yes	No	No	No	No	No	Plausible causal effect
63	Wang, JC; Zhao, YY; Sun, SL; Zhu, JG	Journal Of Business Research	(2023)	Yes	No	No	No	No	No	No	No	Implausible causal effect
64	Zaccane, MC; Argiolas, A	Corporate Governance-The International Journal Of Business In Society	(2023)	Yes	Yes	No	No	No	No	No	No	Implausible causal effect

Appendix B. Overview of coded studies using an instrumental variable design

No.	Authors	Year	Endogenous variables	Instrumental variables	Instrument strength tested	F-value	Internal/ external instrument	Instrument classification	Instrument exogeneity	Endogeneity test	Exclusion restriction
1	Sieweke, J; Bostandzic, D; Smolinski, SM	2023	TMT gender diversity	shift-share instrument	Olea-Pfluger	28.25	external instrument	potentially appropriate	not applicable (only one instrument)	C-test	discussed
2	Khatri, I	2023	board gender diversity	industry average board gender diversity; board gender diversity at the initial period	Cragg-Donald F statistic	84,025	external instruments	potentially appropriate	Hansen test (p = 0.07)	not reported	not discussed

(continued on next page)

(continued)

No.	Authors	Year	Endogenous variables	Instrumental variables	Instrument strength tested	F-value	Internal/external instrument	Instrument classification	Instrument exogeneity	Endogeneity test	Exclusion restriction
3	Havrylyshyn, A; Schepker, DJ; Nyberg, AJ	2023	board gender diversity	industry average women directors (tally); industry average portion of board who are women;	F-statistic	40.60	external instruments	potentially appropriate	Sargan test (p = 0.28)	not reported	partially discussed (authors discuss why the IV is unrelated to the DV)
4	Ferrari, G; Ferraro, V; Profeta, P; Pronzato, C	2022	board gender diversity	Italian board gender quota	F-statistic	119.88	external instruments	potentially appropriate	not applicable	not reported	discussed
5	Wu, J; Richard, OC; Triana, MD; Zhang, XH	2022	board gender diversity; TMT gender diversity	State ownership; TMT size; board size; CEO gender; board chair's gender; industry dummies; year dummies	Cragg-Donald F statistic	10.51	internal instruments	inappropriate	Hansen test (p = 0.00)	p = 0.426	not discussed
6	Yanadori, Y; Kulik, CT; Gould, JA	2021	TMT gender diversity	industry mean of women's representation in the Australian state where the firm's headquarters is located; lagged values of the endogenous variable	not reported		external instrument; internal instrument	potentially appropriate; inappropriate	not reported	not reported	not discussed
7	Fernando, GD; Jain, SS; Tripathy, A	2020	TMT gender diversity	board gender diversity	not reported		internal instrument	inappropriate	not applicable (only one instrument)	not reported	not discussed
8	Wang, Y; Abbasi, K; Babajide, B; Yekini, KC	2019	board gender diversity	lagged endogenous treatment variables	Not reported	Not reported	internal instrument	inappropriate	not reported	not reported	not discussed
9	Martínez, MDV; Rambaud, SC; Oller, IMP	2019	board gender diversity	the number of directors (BSize), the existence of a limited age for directors (Alimit), the existence of a limited managerial period for independent directors (Ylimit), and the existence of a mandatory gender law (Law)	F-test	Not reported (board size);2.75 (age limit);0.52 (year limit) ; 76.06 (gender quota law)	internal instrument; external instruments	inappropriate and potentially appropriate	not reported	not reported	not discussed
10	Giannetti, M; Zhao, MX	2019	board ancestral diversity	ancestral diversity in the county where a firm is headquartered	Kleibergen-Paap Wald test	11.69; 10.39; 11:40	external instrument	potentially appropriate	not applicable	not reported	discussed
11	Delis, MD; Gaganis, C; Hasan, I; Pasiouras, F	2017	genetic diversity	migratory distance; ultraviolet exposure	Wald F-statistic	23,5324,18	external instrument	potentially appropriate	Hansen test (0.06; 0.22)		discussed
12	Conyon, MJ; He, LR	2017	board gender diversity	percentage of female residents in the US state where the given company has its headquarter; alternative instrument: Percentage of women who work in the given company's industry	not reported		external instrument	potentially appropriate	not applicable (only one instrument)	not reported	discussed
13	Adams, RB; Ferreira, D	2009	board gender diversity	fraction of male directors with board connections to female directors	not reported (but can be calculated based on information from table 9, model 3)	11.4921	external instrument	potentially appropriate	not applicable	Hausman test (-2.17)	discussed
14	Uyar, A; Kuzey, C; Kilic, M; Karaman, AS	2021	board gender diversity	lagged independent variables	not reported	not reported	internal instrument	inappropriate	not applicable	not reported	discussed

(continued on next page)

(continued)

No.	Authors	Year	Endogenous variables	Instrumental variables	Instrument strength tested	F-value	Internal/external instrument	Instrument classification	Instrument exogeneity	Endogeneity test	Exclusion restriction
15	Pucheta-Martínez, MC; Bel-Oms, I; Olcina-Sempere, G	2018	female institutional directors	not reported	not reported	not reported	cannot be assessed	cannot be assessed	Hansen test	not reported	not discussed
16	Ahern, KR; Dittmar, AK	2012	board gender diversity	pre-quota board gender ratio	F-value	29.79	internal instrument	potentially appropriate	not applicable	not reported	not discussed

Appendix C. Overview of coded studies using a difference and system GMM

No.	Authors	Year	Estimator	Endogenous variables	Endogeneity test	Number of instruments	Sensitivity test for a reduced number of instruments	Instrument exogeneity	Instrument exogeneity (subset)	Test for Autocorrelation
1	Dezso, CL; Ross, DG	2012	not reported (probably system GMM)	TMT gender diversity	not reported	not reported	not reported	not reported	not reported	not reported
2	Ullah, I; Fang, HX; Jebran, K	2019	GMM (no further information)	board gender diversity	not reported	not reported	not reported	Sargan-Hansen test (72.77)	not reported	Arellano Bond test (AR[2] = -0.72)
3	Mazzotta, R; Ferraro, O	2020	system GMM	board gender diversity	not reported	not reported	not reported	Hansen test (0.379)	not reported	Arellano Bond test (AR[2] p = 0.450)
4	Rehman, S; Orij, R; Khan, H	2020	GMM (no further information)	board gender diversity	not reported	not reported	not reported	Sargan test (p = 0.000)	not reported	Arellano Bond test (AR[2] p = 0.544; 0.862; 0.921)
5	Ali, F; Wang, M; Jebran, K; Ali, ST	2021	system GMM	board gender diversity board age diversity	not reported	not reported	not reported	Sargan test (19.56)	not reported	not reported
6	Saleh, MWA; Zaid, MAA; Shurafa, R; Maigoshi, ZS; Mansour, M; Zaid, A	2021	system GMM	board gender diversity	Durbin-Wu-Hausman	not reported	not reported	Hansen test (0.186; 0.132)	not reported	Arellano Bond test (AR[2] p = 0.277; 0.265)
7	Boukattaya, S; Ftiti, Z; Ben Arfa, N; Omri, A	2022	unclear	board gender diversity		not reported	not reported	Hansen test (0.9520; 0.9495; 0.8880; 0.8672; 0.8892)	not reported	Arellano Bond test (AR[2] p = 0.3177; 0.3686; 0.3278; 0.3382; 0.4332)
8	Akhter, W; Hassan, A	2023	system GMM	board gender diversity	Durbin-Wu-Hausman	not reported	not reported	Sargan test	not reported	Arellano Bond test
9	Ben Fatma, H; Chouaibi, J	2023	system GMM	board gender diversity	not reported	not reported	not reported	Hansen test Sargan test	not reported	Arellano Bond test
10	Fayyaz, UER; Jalal, RNUD; Venditti, M; Minguez-Vera, A	2023	system GMM	board gender diversity board age diversity	not reported	not reported	not reported	Hansen test (p = 0.999; 0.992)	not reported	Arellano Bond test (AR[2] p = 0.117; 0.215)
11	Khatri, I	2023	system GMM	board gender diversity	not reported	not reported	not reported	Hansen test (p = 0.619)	not reported	Arellano Bond test (AR[2] p = 0.08; AR[3] = 0.14)

Appendix D. Overview of coded studies using matching methods

No.	Authors	Year	Distance measure	Matching method	Balance test	Unconfoundedness assumption	Sensitivity analyses
1	Zaccone, MC; Argiolas, A	2023	propensity score	not reported	not reported	not discussed	not reported
2	Rehman, S; Orij, R; Khan, H	2020	propensity score	not reported	not reported	not discussed	not reported

Appendix E. Overview of coded studies using a difference-in-differences design

No.	Authors	Year	Treatment	Treatment group/ control group	Time periods	Treatment adoption	Parallel trend assumption	Strict exogeneity	Standard errors	Matching
1	Yang, P; Riepe, J; Moser, K; Pull, K; Terjesen, S	2019	Norwegian board gender quota	Norwegian firms vs. Firms from Denmark, Sweden and Finland	2002–2008	simultaneous	Tested (graph)	discussed	clustered by firm	one-to-one matching

Data availability

Data will be made available on request.

References

Adams, R. B., & Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics*, 94(2), 291–309. <https://doi.org/10.1016/j.jfineco.2008.10.007>

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-Practice Recommendations for Producers, Evaluators, and Users of Methodological Literature Reviews. *Organizational Research Methods*, 26(1), 46–76. <https://doi.org/10.1177/1094428120943281>

Ahern, K. R., & Dittmar, A. K. (2012). The Changing of the Boards: The Impact on Firm Valuation of Mandated Female Board Representation. *The Quarterly Journal of Economics*, 127(1), 137–197. <https://doi.org/10.1093/qje/qjr049>

Akhter, W., & Hassan, A. (2023). Does corporate social responsibility mediate the relationship between corporate governance and firm performance? Empirical evidence from BRICS countries. *Corporate Social Responsibility and Environmental Management*, 31(1), 566–578. <https://doi.org/10.1002/csr.2586>

Aldhamari, R., Mohamad Nor, M. N., Boudiab, M., & Mas'ud, A. (2020). The impact of political connection and risk committee on corporate financial performance: Evidence from financial firms in Malaysia. *Corporate Governance: The International Journal of Business in Society*, 20(7), 1281–1305. <https://doi.org/10.1108/cg-04-2020-0122>

Ali, F., Wang, M., Jebran, K., & Ali, S. T. (2021). Board diversity and firm efficiency: Evidence from China. *Corporate Governance: The International Journal of Business in Society*, 21(4), 587–607. <https://doi.org/10.1108/cg-10-2019-0312>

Alodat, A. Y., Salleh, Z., Nobanee, H., & Hashim, H. A. (2023). Board gender diversity and firm performance: The mediating role of sustainability disclosure. *Corporate Social Responsibility and Environmental Management*, 30(4), 2053–2065. <https://doi.org/10.1002/csr.2473>

Andoh, J. A. N., Abugri, B. A., & Anarfo, E. B. (2023). Board Characteristics and performance of listed firms in Ghana. *Corporate Governance: The International Journal of Business in Society*, 23(1), 43–71. <https://doi.org/10.1108/cg-08-2020-0344>

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1), 727–753. <https://doi.org/10.1146/annurev-economics-080218-025643>

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, 28(1), 5–21. <https://doi.org/10.1016/j.leaqua.2017.01.006>

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086–1120. <https://doi.org/10.1016/j.leaqua.2010.10.010>

Arena, C., Cirillo, A., Mussolino, D., Pulcinelli, I., Saggese, S., & Sarto, F. (2015). Women on board: Evidence from a masculine industry. *Corporate Governance*, 15(3), 339–356. <https://doi.org/10.1108/cg-02-2014-0015>

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3–32. <https://doi.org/10.1257/jep.31.2.3>

Bastardoz, N., Jacquart, P., & Antonakis, J. (2024). Effect of crises on charisma signaling: A regression discontinuity design. *The Leadership Quarterly*, 101590. <https://doi.org/10.1016/j.leaqua.2021.101590>

Bastardoz, N., Matthews, M. J., Sajons, G. B., Ransom, T., Kelemen, T. K., & Matthews, S. H. (2023). Instrumental variables estimation: Assumptions, pitfalls, and guidelines. *The Leadership Quarterly*, 34(1), Article 101673. <https://doi.org/10.1016/j.leaqua.2022.101673>

Ben-Shahar, D., Carmeli, A., Sulganik, E., & Weiss, D. (2024). Regulating Diversity and Inclusiveness in Boards of Directors. *Academy of Management Perspectives*, 38(3), 456–461. <https://doi.org/10.5465/amp.2023.0474>

Ben Fatma, H., & Chouaibi, J. (2023). Gender diversity, financial performance, and the moderating effect of CSR: Empirical evidence from UK financial institutions. *Corporate Governance: The International Journal of Business in Society*, 23(7), 1506–1525. <https://doi.org/10.1108/cg-11-2022-0445>

Boukattaya, S., Ftiti, Z., Ben Arfa, N., & Omri, A. (2022). Financial performance under board gender diversity: The mediating effect of corporate social practices. *Corporate Social Responsibility and Environmental Management*, 29(5), 1871–1883. <https://doi.org/10.1002/csr.2333>

Calabrese, G. G., & Manello, A. (2021). Board diversity and performance in a masculine, aged and global supply chain: New empirical evidence. *Corporate Governance: The International Journal of Business in Society*, 21(7), 1440–1459. <https://doi.org/10.1108/cg-09-2020-0417>

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. (2024). *Difference-in-differences with a continuous treatment* (32117). National Bureau of Economic Research. <https://www.nber.org/papers/w32117>

Campbell, K., & Mínguez-Vera, A. (2007). Gender Diversity in the Boardroom and Firm Financial Performance. *Journal of Business Ethics*, 83(3), 435–451. <https://doi.org/10.1007/s10551-007-9630-y>

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs*: (Volume 1). Cambridge University Press.

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, 115(531), 1449–1455. <https://doi.org/10.1080/01621459.2019.1635480>

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, 14, 821–851. <https://doi.org/10.1146/annurev-economics-051520-021409>

Celli, V. (2022). Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys*, 36(1), 214–234. <https://doi.org/10.1111/joes.12452>

Certo, S. T., Busenbark, J. R., Woo, H.s., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, 37(13), 2639–2657.

Chong, L.-L., Ong, H.-B., & Tan, S.-H. (2018). Corporate risk-taking and performance in Malaysia: The effect of board composition, political connections and sustainability practices. *Corporate Governance: The International Journal of Business in Society*, 18(4), 635–654. <https://doi.org/10.1108/cg-05-2017-0095>

Conley, T. G., Hansen, C. B., & Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94(1), 260–272. https://doi.org/10.1162/REST_a_00139

Conyon, M. J., & He, L. (2017). Firm performance and boardroom gender diversity: A quantile regression approach. *Journal of Business Research*, 79, 198–211. <https://doi.org/10.1016/j.jbusres.2017.02.006>

Cronin, M. A., & George, E. (2023). The Why and How of the Integrative Review. *Organizational Research Methods*, 26(1), 168–192. <https://doi.org/10.1177/1094428120935507>

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and Big Data Are Changing Economics. *American Economic Review Papers and Proceedings*, 110, 42–48. <https://doi.org/10.1257/pandp.20201058>

Darko, J., Aribi, Z. A., & Uzonwanne, G. C. (2016). Corporate governance: The impact of director and board structure, ownership structure and corporate control on the performance of listed companies on the Ghana stock exchange. *Corporate Governance*, 16(2), 259–277. <https://doi.org/10.1108/cg-11-2014-0133>

Darmadi, S. (2013). Do women in top management affect firm performance? Evidence from Indonesia. *Corporate Governance: The International Journal of Business in Society*, 13(3), 288–304. <https://doi.org/10.1108/cg-12-2010-0096>

Dehejia, R. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics*, 125(1–2), 355–364. <https://doi.org/10.1016/j.jeconom.2004.04.012>

Delis, M. D., Gaganis, C., Hasan, I., & Pasiouras, F. (2017). The effect of board directors from countries with different genetic diversity levels on corporate performance. *Management Science*, 63(1), 231–249. <https://doi.org/10.1287/mnsc.2015.2299>

Der Bundesrat. Das Portal der Schweizer Regierung. (2020). *Geschlechterschritte und Transparenzregeln für Rohstoffsektor treten Anfang 2021 in Kraft*. <https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-80358.html>

Dezso, C. L., & Ross, D. G. (2012). Does female representation in top management improve firm performance? A panel data investigation. *Strategic Management Journal*, 33(9), 1072–1089. <https://doi.org/10.1002/smj.1955>

Dwyer, S., Richard, O. C., & Chadwick, K. (2003). Gender diversity in management and firm performance: The influence of growth orientation and organizational culture. *Journal of Business Research*, 56(12), 1009–1019. [https://doi.org/10.1016/s0148-2963\(01\)00329-0](https://doi.org/10.1016/s0148-2963(01)00329-0)

Erhardt, N. L., Werbel, J. D., & Shrader, C. B. (2003). Board of director diversity and firm financial performance. *Corporate Governance*, 11(2), 102–111. <https://doi.org/10.1111/1467-8683.00011>

Farooq, M., & Ahmad, N. (2023). Nexus between board characteristics, firm performance and intellectual capital: An emerging market evidence. *Corporate Governance: The International Journal of Business in Society*, 23(6), 1269–1297. <https://doi.org/10.1108/cg-08-2022-0355>

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

- disclosure. *Corporate Social Responsibility and Environmental Management*, 30(3), 1457–1472. <https://doi.org/10.1002/csr.2430>
- Fernández-Temprano, M. A., & Tejerina-Gaite, F. (2020). Types of director, board diversity and firm performance. *Corporate Governance: The International Journal of Business in Society*, 20(2), 324–342. <https://doi.org/10.1108/cg-03-2019-0096>
- Fernando, G. D., Jain, S. S., & Tripathy, A. (2020). This cloud has a silver lining: Gender diversity, managerial ability, and firm performance. *Journal of Business Research*, 117, 484–496. <https://doi.org/10.1016/j.jbusres.2020.05.042>
- Ferrari, G., Ferraro, V., Profeta, P., & Pronzato, C. (2022). Do Board Gender Quotas Matter? Selection, Performance, and Stock Market Effects. *Management Science*, 68(8), 5618–5643. <https://doi.org/10.1287/mnsc.2021.4200>
- Foster, B. P., Manikas, A. S., & Kroes, J. R. (2023). Which diversity measures best capture public company value? *Corporate Social Responsibility and Environmental Management*, 30(1), 236–247. <https://doi.org/10.1002/csr.2351>
- Francoeur, C., Labelle, R., & Sinclair-Desgagné, B. (2007). Gender Diversity in Corporate Governance and Top Management. *Journal of Business Ethics*, 81(1), 83–95. <https://doi.org/10.1007/s10551-007-9482-5>
- Gharbi, S., & Othmani, H. (2023). Threshold effects of board gender diversity on firm performance: Panel smooth transition regression model. *Corporate Governance: The International Journal of Business in Society*, 23(1), 243–261. <https://doi.org/10.1108/cg-10-2021-0373>
- Giannetti, M., & Zhao, M. (2019). Board ancestral diversity and firm performance volatility. *Journal of Financial and Quantitative Analysis*, 54(3), 1117–1155. <https://doi.org/10.1017/S0022109018001023>
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277. <https://doi.org/10.3386/w25018>
- Hamann, P. M., Schiemann, F., Bellora, L., & Guenther, T. W. (2013). Exploring the dimensions of organizational performance: A construct validity study. *Organizational Research Methods*, 16(1), 67–87. <https://doi.org/10.1177/1094428112470007>
- Hassell, H. J., & Holbein, J. B. (2024). Navigating potential pitfalls in difference-in-differences designs: Reconciling conflicting findings on mass shootings' effect on electoral outcomes. *American Political Science Review*, 1–21. <https://doi.org/10.1017/s0003055424000108>
- Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10, 533–552. <https://doi.org/10.3386/w23602>
- Havrylyshyn, A., Schepker, D. J., & Nyberg, A. J. (2023). In the Club? How Categorization and Contact Impact the Board Gender Diversity-Firm Performance Relationship. *Journal of Business Ethics*, 184(2), 353–374. <https://doi.org/10.1007/s10551-022-05168-0>
- Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review*, 74(2), 208–224. <https://doi.org/10.1177/000312240907400203>
- Hill, T. D., Davis, A. P., Roos, J. M., & French, M. T. (2020). Limitations of fixed-effects models for panel data. *Sociological Perspectives*, 63(3), 357–369. <https://doi.org/10.1177/0731121419863785>
- Hoobler, J. M., Masterson, C. R., Nkomo, S. M., & Michel, E. J. (2018). The business case for women leaders: Meta-analysis, research critique, and path forward. *Journal of Management*, 44(6), 2473–2499. <https://doi.org/10.5703/1288284316077>
- Iacus, S. M., King, G., & Porro, G. (2019). A theory of statistical inference for matching methods in causal research. *Political Analysis*, 27(1), 46–68. <https://doi.org/10.1017/pan.2018.29>
- Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of applied econometrics*, 23(3), 305–327. <https://doi.org/10.1002/jae.998>
- Imai, K., Kim, I. S., & Wang, E. H. (2023). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3), 587–605. <https://doi.org/10.1111/ajps.12685>
- Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2), 373–419. <https://doi.org/10.3386/w19959>
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86. <https://doi.org/10.1257/jel.47.1.5>
- Isidro, H., & Sobral, M. (2014). The Effects of Women on Corporate Boards on Firm Value, Financial Performance, and Ethical and Social Compliance. *Journal of Business Ethics*, 132(1), 1–19. <https://doi.org/10.1007/s10551-014-2302-9>
- Jacquart, P., Santoni, S., Schudy, S., Sieweke, J., & Withers, M. (2024). Exogenous shocks: Definitions, types, and causal identification issues. *The Leadership Quarterly*, 35(5), Article 101823.
- Joecks, J., Pull, K., & Vetter, K. (2012). Gender Diversity in the Boardroom and Firm Performance: What Exactly Constitutes a “Critical Mass?”. *Journal of Business Ethics*, 118(1), 61–72. <https://doi.org/10.1007/s10551-012-1553-6>
- Kahn-Lang, A., & Lang, K. (2020). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38(3), 613–620. <https://doi.org/10.3386/w24857>
- Khatri, I. (2023). Board gender diversity and sustainability performance: Nordic evidence. *Corporate Social Responsibility and Environmental Management*, 30(3), 1495–1507. <https://doi.org/10.1002/csr.2432>
- Kirsch, A. (2018). The gender composition of corporate boards: A review and research agenda. *The Leadership Quarterly*, 29(2), 346–364. <https://doi.org/10.1016/j.leaqua.2017.06.001>
- Krause, R., Roh, J., & Whitley, K. A. (2022). The top management team: Conceptualization, operationalization, and a roadmap for scholarship. *Journal of Management*, 48(6), 1548–1601. <https://doi.org/10.1177/01492063211072459>
- Lal, A., Lockhart, M., Xu, Y., & Zu, Z. (2023). How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies. *Political Analysis*, 1–20. <https://doi.org/10.1017/pan.2024.2>
- Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics*, 49(3), 186–205. <https://doi.org/10.1016/j.jacceco.2009.11.004>
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281–355. <https://doi.org/10.1257/jel.48.2.281>
- Li, J., Ding, H., Hu, Y., & Wan, G. (2021). Dealing with dynamic endogeneity in international business research. *Journal of International Business Studies*, 52, 339–362. <https://doi.org/10.1057/s41267-020-00398-8>
- Liu, Y., Lei, L., & Buttner, E. H. (2020). Establishing the boundary conditions for female board directors' influence on firm performance through CSR. *Journal of Business Research*, 121, 112–120. <https://doi.org/10.1016/j.jbusres.2020.08.026>
- Luciano, M. M., Nahrgang, J. D., & Shropshire, C. (2020). Strategic leadership systems: Viewing top management teams and boards of directors from a multiteam systems perspective. *Academy of Management Review*, 45(3), 675–701. <https://doi.org/10.5465/amr.2017.0485>
- Martinez-Jimenez, R., Hernández-Ortiz, M. J., & Cabrera Fernández, A. I. (2020). Gender diversity influence on board effectiveness and business performance. *Corporate Governance: The International Journal of Business in Society*, 20(2), 307–323. <https://doi.org/10.1108/cg-07-2019-0206>
- Matsa, D. A., & Miller, A. R. (2013). A female style in corporate leadership? Evidence from quotas. *American Economic Journal: Applied Economics*, 5(3), 136–169. <https://doi.org/10.1257/app.5.3.136>
- Mazzotta, R., & Ferraro, O. (2020). Does the gender quota law affect bank performances? Evidence from Italy. *Corporate Governance: The International Journal of Business in Society*, 20(6), 1135–1158. <https://doi.org/10.1108/cg-08-2019-0252>
- McGuinness, P. B. (2018). IPO Firm Performance and Its Link with Board Officer Gender, Family-Ties and Other Demographics. *Journal of Business Ethics*, 152(2), 499–521. <https://doi.org/10.1007/s10551-016-3295-3>
- McIntyre, M. L., Murphy, S. A., & Mitchell, P. (2007). The top team: Examining board composition and firm performance. *Corporate Governance: The International Journal of Business in Society*, 7(5), 547–561. <https://doi.org/10.1108/14720700710827149>
- Montiel Olea, J. L., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3), 358–369. <https://doi.org/10.1080/00401706.2013.806694>
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.
- Morgenroth, T., & Ryan, M. K. (2018). Quotas and affirmative action: Understanding group-based outcomes and attitudes. *Social and Personality Psychology Compass*, 12(3), e12374.
- Narita, K., Tena, J., & Detotto, C. (2023). Causal inference with observational data: A tutorial on propensity score analysis. *The Leadership Quarterly*, 101678. <https://doi.org/10.1016/j.leaqua.2023.101678>
- Neely, B. H., Lovelace, J. B., Cowen, A. P., & Hiller, N. J. (2020). Metacritiques of upper echelons theory: Verdicts and recommendations for future research. *Journal of Management*, 46(6), 1029–1062. <https://doi.org/10.1177/01492063200908640>
- Nielsen, B. B., & Nielsen, S. (2012). Top management team nationality diversity and firm performance: A multilevel study. *Strategic Management Journal*, 34(3), 373–382. <https://doi.org/10.1002/smj.2021>
- O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, 16(1), 1–21. <https://doi.org/10.1007/s10742-016-0146-8>
- Olden, A., & Moen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3), 531–553. <https://doi.org/10.1093/ectj/utac010>
- Perryman, A. A., Fernando, G. D., & Tripathy, A. (2016). Do gender differences persist? An examination of gender diversity on firm performance, risk, and executive compensation. *Journal of Business Research*, 69(2), 579–586. <https://doi.org/10.1016/j.jbusres.2015.05.013>
- Pucheta-Martínez, M. C., Bel-Oms, I., & Olcina-Sempere, G. (2018). Female Institutional Directors on Boards and Firm Value. *Journal of Business Ethics*, 152(2), 343–363. <https://doi.org/10.1007/s10551-016-3265-9>
- Rambachan, A., & Roth, J. (2023). An honest approach to parallel trends. *Review of Economic Studies*, 90, 2555–2591. <https://doi.org/10.1093/restud/rdad018>
- Rehman, S., Orij, R., & Khan, H. (2020). The search for alignment of board gender diversity, the adoption of environmental management systems, and the association with firm performance in Asian firms. *Corporate Social Responsibility and Environmental Management*, 27(5), 2161–2175. <https://doi.org/10.1002/csr.1955>
- Roberson, Q. M., Holmes, O., IV, & Perry, J. L. (2017). Transforming research on diversity and firm performance: A dynamic capabilities perspective. *Academy of Management Annals*, 11(1), 189–216. <https://doi.org/10.5465/annals.2014.0019>
- Roodman, D. (2009a). How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal*, 9(1), 86–136. <https://doi.org/10.1177/1536867X0900900106>
- Roodman, D. (2009b). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1), 135–158. <https://doi.org/10.1111/j.1468-0084.2008.00542.x>
- Rose, C. (2007). Does female board representation influence firm performance? The Danish evidence. *Corporate Governance: An International Review*, 15(2), 404–413. <https://doi.org/10.1111/j.1467-8683.2007.00570.x>
- Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3), 305–322. <https://doi.org/10.1257/aeri.20210236>

- Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2), 2218–2244. <https://doi.org/10.1016/j.jeconom.2023.03.008>
- Roudaki, J. (2018). Corporate governance structures and firm performance in large agriculture companies in New Zealand. *Corporate Governance: The International Journal of Business in Society*, 18(5), 987–1006. <https://doi.org/10.1108/cg-07-2018-0241>
- Saleh, M. W. A., Zaid, M. A. A., Shurafa, R., Maigoshi, Z. S., Mansour, M., & Zaid, A. (2021). Does board gender enhance Palestinian firm performance? The moderating role of corporate social responsibility. *Corporate Governance: The International Journal of Business in Society*, 21(4), 685–701. <https://doi.org/10.1108/cg-08-2020-0325>
- Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12, 487–508. <https://doi.org/10.1146/annurev.polisci.11.060606.135444>
- Semadeni, M., Withers, M. C., & Certo, S. T. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, 35(7), 1070–1079. <https://doi.org/10.1002/smj.2136>
- Sieweke, J., Bostandzic, D., & Smolinski, S.-M. (2023). The influence of top management team gender diversity on firm performance during stable periods and economic crises: An instrumental variable analysis. *The Leadership Quarterly*, 34(5). <https://doi.org/10.1016/j.leaqua.2023.101703>
- Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *The Leadership Quarterly*, 31(1), Article 101338. <https://doi.org/10.1016/j.leaqua.2019.101338>
- Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2), 305–353. <https://doi.org/10.1016/j.jeconom.2004.04.011>
- State of California. (2018). *Senate Bill No. 826*. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB826
- State of California. (2020). *Assembly Bill No. 979*. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB979
- Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 237–259). Oxford University Press.
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews, & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1), 1–21. <https://doi.org/10.1214/09-STS313>
- Toumi, N., Benkraiem, R., & Hamrouni, A. (2016). Board director disciplinary and cognitive influence on corporate value creation. *Corporate Governance*, 16(3), 564–578. <https://doi.org/10.1108/cg-09-2015-0123>
- Triana, M. d. C., Richard, O. C., & Su, W. (2019). Gender diversity in senior management, strategic change, and firm performance: Examining the mediating nature of strategic change in high tech firms. *Research Policy*, 48(7), 1681–1693. <https://doi.org/10.1016/j.respol.2019.03.013>
- Ujunwa, A. (2012). Board characteristics and the financial performance of Nigerian quoted firms. *Corporate Governance: The International Journal of Business in Society*, 12(5), 656–674. <https://doi.org/10.1108/14720701211275587>
- Ullah, I., Fang, H., & Jebran, K. (2019). Do gender diversity and CEO gender enhance firm's value? Evidence from an emerging economy. *Corporate Governance: The International Journal of Business in Society*, 20(1), 44–66. <https://doi.org/10.1108/cg-03-2019-0085>
- Ullah, S., Akhtar, P., & Zaeferian, G. (2018). Dealing with endogeneity bias: The generalized method of moments (GMM) for panel data. *Industrial Marketing Management*, 71, 69–78. <https://doi.org/10.1016/j.indmarman.2017.11.010>
- Uribe-Bohorquez, M. V., Martínez-Ferrero, J., & García-Sánchez, I. M. (2019). Women on boards and efficiency in a business-orientated environment. *Corporate Social Responsibility and Environmental Management*, 26(1), 82–96. <https://doi.org/10.1002/csr.1659>
- Uyar, A., Kuzey, C., Kilic, M., & Karaman, A. S. (2021). Board structure, financial performance, corporate social responsibility performance, CSR committee, and CEO duality: Disentangling the connection in healthcare. *Corporate Social Responsibility and Environmental Management*, 28(6), 1730–1748. <https://doi.org/10.1002/csr.2141>
- Vairavan, A., & Zhang, G. P. (2020). Does a diverse board matter? A mediation analysis of board racial diversity and firm performance. *Corporate Governance: The International Journal of Business in Society*, 20(7), 1223–1241. <https://doi.org/10.1108/cg-02-2020-0081>
- Van der Walt, N., Ingleby, C., Shergill, G. S., & Townsend, A. (2006). Board configuration: Are diverse boards better boards? *Corporate Governance: The International Journal of Business in Society*, 6(2), 129–147. <https://doi.org/10.1108/14720700610655141>
- Van Doorn, S., Heyden, M. L. M., & Reimer, M. (2023). The private life of CEOs: A strategic leadership perspective. *The Leadership Quarterly*, 34(1), Article 101679. <https://doi.org/10.1016/j.leaqua.2023.101679>
- Veltri, S., Mazzotta, R., & Rubino, F. E. (2021). Board diversity and corporate social performance: Does the family firm status matter? *Corporate Social Responsibility and Environmental Management*, 28(6), 1664–1679. <https://doi.org/10.1002/csr.2136>
- Vera, D., Bonardi, J.-P., Hitt, M. A., & Withers, M. C. (2022). Extending the boundaries of strategic leadership research. *The Leadership Quarterly*, 33(3), Article 101617. <https://doi.org/10.1016/j.leaqua.2022.101617>
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Wang, J. C., Zhao, Y., Sun, S. L., & Zhu, J. (2023). Female-friendly boards in family firms. *Journal of Business Research*, 157. <https://doi.org/10.1016/j.jbusres.2022.113552>
- Wang, Y., Abbasi, K., Babajide, B., & Yekini, K. C. (2019). Corporate governance mechanisms and firm performance: Evidence from the emerging market following the revised CG code. *Corporate Governance: The International Journal of Business in Society*, 20(1), 158–174. <https://doi.org/10.1108/cg-07-2018-0244>
- Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39(1), 453–469. <https://doi.org/10.1146/annurev-publhealth-040617-013507>
- Wooldridge, J. M. (2009). *Introductory econometrics - A modern approach*. Thompson Higher Education.
- Wu, J., Richard, O. C., Triana, M. d. C., & Zhang, X. (2022). The performance impact of gender diversity in the top management team and board of directors: A multiteam systems approach. *Human Resource Management*, 61(2), 157–180. <https://doi.org/10.1002/hrm.22086>
- Wulff, J. N., Sajons, G. B., Pogrebna, G., Lonati, S., Bastardoz, N., Banks, G. C., & Antonakis, J. (2023). Common methodological mistakes. *The Leadership Quarterly*, 34(1), Article 101677. <https://www.sciencedirect.com/science/article/pii/S1048984323000036>
- Yanadori, Y., Kulik, C. T., & Gould, J. A. (2021). Who pays the penalty? Implications of gender pay disparities within top management teams for firm performance. *Human Resource Management*, 60(4), 681–699. <https://doi.org/10.1002/hrm.22067>
- Yang, P., Riepe, J., Moser, K., Pull, K., & Terjesen, S. (2019). Women directors, firm performance, and firm risk: A causal perspective. *The Leadership Quarterly*, 30(5), Article 101297. <https://doi.org/10.1016/j.leaqua.2019.05.004>
- Zaccone, M. C., & Argiolas, A. (2023). Is a critical mass of women always enough to improve firm performance? The importance of the institutional context. *Corporate Governance: The International Journal of Business in Society*, 24(8), 1–21. <https://doi.org/10.1108/cg-02-2023-0058>
- Zhang, J. Q., Zhu, H., & Ding, H.-B. (2012). Board Composition and Corporate Social Responsibility: An Empirical Investigation in the Post Sarbanes-Oxley Era. *Journal of Business Ethics*, 114(3), 381–392. <https://doi.org/10.1007/s10551-012-1352-0>
- Zhang, L. (2012). Board demographic diversity, independence, and corporate social performance. *Corporate Governance: The International Journal of Business in Society*, 12(5), 686–700. <https://doi.org/10.1108/14720701211275604>