Gender differences in science

Methodological approaches to address common challenges in bibliometric analyses

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Sozialwissenschaftlichen Fakultät

der Ludwig-Maximilians-Universität München

vorgelegt von

Alexander Tekles

2023

Erstgutachterin: Prof. Dr. Katrin Auspurg (Ludwig-Maximilians-Universität München)

Zweitgutachter: Prof. Dr. Thomas Hinz (Universität Konstanz)

Tag der mündlichen Prüfung: 22.05.2023

Danksagung

Für die maßgebliche Unterstützung beim Verfassen dieser Dissertation möchte ich mich besonders herzlich bedanken bei:

- Katrin Auspurg für das überaus wertvolle Feedback und das äußerst angenehme Arbeitsumfeld
- Lutz Bornmann für die ausgezeichnete Zusammenarbeit und die großartige Unterstützung, sowohl fachlich als auch persönlich
- Thomas Hinz für die bereitwillige Übernahme der Rolle als Zweitgutachter
- Thomas Augustin für die Bereitschaft als Drittprüfer zur Verfügung zu stehen
- zahlreichen Kolleg*innen am Institut für Soziologie für den anregenden Austausch und die große Hilfsbereitschaft
- den Kolleg*innen in der Generalverwaltung der Max-Planck-Gesellschaft für die Möglichkeit, neben dem üblichen Wissenschaftsbetrieb wertvolle Eindrücke zu sammeln
- Robin Haunschild für den unkomplizierten Zugriff auf IT-Ressourcen, ohne die einige der Analysen in dieser Arbeit nicht möglich gewesen wären

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Wissenschaftliches Wissen wird meist als eine besondere Form von Wissen betrachtet, die mehr Vertrauen genießt als andere Wissensformen. Dieses Vertrauen wird mitunter dadurch gerechtfertigt, dass Wissenschaft sich durch eine meritokratische Ordnung auszeichnet. Demnach bemisst sich das Ansehen von Wissenschaftler*innen und die Anerkennung von deren Leistung lediglich nach dem Beitrag zum wissenschaftlichen Fortschritt. Merton hat diesen Imperativ bereits in seinem Ethos der Wissenschaft unter der Norm des Universalismus zusammengefasst. Für die Evaluierung und demnach auch die Anerkennung von wissenschaftlichen Leistungen werden u. a. bibliometrische Indikatoren verwendet. Dies setzt voraus, dass bibliometrische Indikatoren in erster Linie die Leistung von Wissenschaftler*innen widerspiegeln und nicht durch andere Faktoren beeinflusst werden.

Insbesondere sollte demnach das Geschlecht der Wissenschaftler*innen keinen Einfluss auf bibliometrische Indikatoren haben. Allerdings deuten viele empirische Ergebnisse darauf hin, dass ein solcher Zusammenhang besteht. Nachdem bibliometrische Indikatoren auch eine wichtige Rolle für wissenschaftliche Karrieren spielen, können entsprechende Geschlechterunterschiede darüber hinaus als möglicher Grund für den nach wie vor geringeren Anteil von Wissenschaftlerinnen im Vergleich zu Wissenschaftlern gesehen werden. Somit ist ein gutes Verständnis des Zusammenhangs zwischen dem Geschlecht von Wissenschaftler*innen und bibliometrischen Indikatoren nötig, um Geschlechterunterschiede in der Wissenschaft im Allgemeinen zu verstehen und geeignete Maßnahmen für deren Verringerung zu ergreifen.

Die vorliegende Dissertation beschäftigt sich mit der Analyse des Zusammenhangs zwischen Geschlecht und bibliometrischen Indikatoren, wobei ein besonderer Fokus auf den methodischen Verfahren zur Datenanalyse liegt. Neben einem einführenden Kapitel besteht die Dissertation aus vier Studien. Die Studie in Kapitel 2 geht der Frage nach, ob sich eine Tendenz zu Geschlechterhomophilie in Zitationsentscheidungen feststellen lässt, d. h. ob Wissenschaftler*innen Personen desselben Geschlechts häufiger zitieren als zu erwarten wäre. In früheren bibliometrischen Analysen konnte ein entsprechendes Muster in Zitationen festgestellt werden. Eine Geschlechterhomophilie in Zitationsentscheidungen könnte zu Geschlechterunterschieden im Citation Impact führen, da insgesamt mehr Männer als Frauen in der Wissenschaft aktiv sind.

Die Studie in Kapitel 5 untersucht Geschlechterunterschiede in bibliometrischen Indikatoren, die den Output von Wissenschaftler*innen messen. Bei den bibliometrischen Indikatoren, die in der Studie betrachtet werden, handelt es sich um die Anzahl der Publikationen, die eine Person veröffentlicht hat, den Citation Impact der Publikationen sowie dem Impact Factor der Zeitschriften, in denen die Publikationen erschienen sind. Diese Indikatoren spielen für Evaluationen von Wissenschaftler*innen und damit auch deren Karriere eine wichtige Rolle. Die Analysen in dieser Studie beziehen sich auf die Personenebene, wohingegen sich frühere bibliometrische Studien meist auf die Publikationsebene beziehen.

Für bibliometrische Analysen auf der Personenebene müssen auch in den verwendeten Daten einzelne Personen repräsentiert sein. Da bibliometrische Daten zunächst nur auf der Publikationsebene vorliegen, sind Verfahren zur Identifikation von Personen nötig. Es muss also bestimmt werden, welche Publikationen zur selben Person und welche Publikationen zu unterschiedlichen Personen gehören. Das grundsätzliche Problem hierbei ist, dass unter den Namen der Autor*innen sowohl Synonyme (unterschiedliche Namen bzw. Schreibweisen für eine Person) als auch Homonyme (identische Namen für unterschiedliche Personen) vorkommen. Zur Auflösung dieser Ambiguitäten wurden mehrere Verfahren vorgeschlagen. Die Studie in Kapitel 4 vergleicht und evaluiert vier dieser Verfahren. Ein Vergleich der Verfahren auf Basis früherer Analysen ist nicht möglich, da die Verfahren anhand unterschiedlicher Daten evaluiert wurden. In der Studie in Kapitel 4 werden die Verfahren unter gleichbleibenden Bedingungen evaluiert, wodurch ein direkter Vergleich möglich ist. Auf Basis der Studie sind auch Rückschlüsse möglich hinsichtlich der Qualität der Daten, die in der Studie in Kapitel 5 zur Analyse auf Personenebene verwendet wurden.

Ein zentraler Aspekt der Studien in den Kapiteln 2 und 5 ist die Kontrolle der thematischen Ausrichtung der Publikationen bzw. der Disziplinen, in denen die Wissenschaftler*innen publiziert haben. Dies ist nötig, da eine geschlechtsspezifische Segregation hinsichtlich wissenschaftlicher Disziplinen zu einem stark variierenden Geschlechterverhältnis zwischen verschiedene Disziplinen führt. Unter der Annahme, dass Zitationen überwiegend innerhalb von Disziplinen erfolgen, führt dieser Umstand dazu, dass Wissenschaftler*innen häufiger von anderen Personen desselben Geschlechts zitiert werden. Dies ist selbst bei einer rein zufälligen Verteilung der Zitationen innerhalb der Disziplinen der Fall. Insofern darf dieses Zitationsmuster in der Studie in Kapitel 2 nicht mit einer Geschlechterhomophilie in Zitationsentscheidungen verwechselt werden, was durch die Kontrolle der thematischen Ausrichtung der Publikationen ermöglicht wird.

Da sich verschiedene Disziplinen ferner hinsichtlich der vorherrschenden Zitations- und Publikationskulturen unterscheiden, kann die geschlechtsspezifische Segregation in verschiedene Disziplinen auch zu Geschlechterunterschieden in den Indikatoren führen, die in der Studie in Kapitel 5 betrachtet werden. Um diese Unterschiede von anderen Mechanismen wie einer geschlechtsbezogenen Diskriminierung in der Bewertung wissenschaftlicher Leistung zu trennen, werden auch in dieser Studie Disziplinen in den Analysen berücksichtigt. Da die Analysen auf Personenebene erfolgen, werden hierfür die Disziplinen verwendet, in denen die Wissenschaftler*innen im Laufe ihrer Karriere publiziert haben.

Für die Kontrolle der thematischen Ausrichtung der Publikationen bzw. der Disziplinen, in denen die Wissenschaftler*innen publiziert haben, wird in den beiden Studien ein ähnlicher Ansatz verfolgt. In früheren Studien wurden zu diesem Zweck Publikationen bzw. Wissenschaftler*innen zu einzelnen Disziplinen zugeordnet. Für die Analysen der hier vorliegenden Studien wurden dagegen Ähnlichkeiten zwischen Publikationen bzw. Wissenschaftler*innen hinsichtlich ihrer thematischen Ausrichtung bzw. der Disziplinen, in denen Wissenschaftler*innen

publiziert haben, berechnet. Durch die Konzentration der Analysen auf Paare von ähnlichen Publikationen bzw. Wissenschaftler*innen ist eine Kontrolle der thematischen Ausrichtung bzw. Disziplinen möglich. Dieser Ansatz hat gegenüber herkömmlichen Verfahren zur Kontrolle der geschlechtsspezifischen Segregation in verschiedene Disziplinen den Vorteil, dass die relevanten Unterschiede zwischen Disziplinen flexibler und genauer erfasst werden können.

Der Ansatz Paare von ähnlichen Publikationen bzw. Wissenschaftler*innen zu identifizieren entspricht der grundsätzlichen Vorgehensweise bei Matching-Verfahren. Die Studie in Kapitel 3 gibt einen Überblick über verschiedene solcher Matching-Verfahren. Die Studie zielt darauf ab, die Verwendung dieser Verfahren in der Szientometrie zu fördern. Dafür werden in der Studie ausgewählte Verfahren vorgestellt und exemplarisch auf bibliometrische Daten angewendet.

Die Ergebnisse der Analysen in den Kapiteln 2 und 5 bestätigen die Notwendigkeit einer sorgfältigen Kontrolle der geschlechtsspezifischen Segregation in verschiedene Disziplinen. Bei einer hinreichenden Kontrolle der thematischen Ausrichtung der Publikationen lassen sich nahezu keine Anzeichen für eine Geschlechterhomophilie bei Zitationsentscheidungen feststellen.

Auch die Geschlechterunterschiede in den bibliometrischen Indikatoren verringern sich, sobald die Disziplinen, in denen die Wissenschaftler*innen publiziert haben, kontrolliert werden. Dennoch lassen sich noch Geschlechterunterschiede in den Indikatoren beobachten. Während Wissenschaftler mehr Publikationen haben, erreichen die Publikationen von Wissenschaftlerinnen einen höheren Citation Impact und erscheinen in Zeitschriften mit höherem Impact Factor. Diese Ergebnisse deuten darauf hin, dass sich die Auswahl bibliometrischer Indikatoren unterschiedlich auf die Evaluation von Wissenschaftlerinnen und Wissenschaftlern auswirkt. Durch eine weitere Differenzierung der Ergebnisse nach Karrierelänge zeigt sich außerdem, dass insbesondere Frauen die Wissenschaft verlassen, deren Publikationen einen hohen Citation Impact haben und die in Zeitschriften mit hohem Impact Factor publizieren.

Beiträge zu den Artikeln

Folgende Tabelle listet alle Artikel auf, die Bestandteil dieser Dissertation sind. Dargestellt sind die Beiträge in Form von CRediT (Contributor Roles Taxonomy) Kategorien sowie der Gesamtanteil an den Artikeln. Die Beiträge zu der Studie in Kapitel 3 beziehen sich insbesondere auf die Erzeugung des Datensatzes, die Beschreibung der Indikatoren (Kapitel 3.2) und die Analysen mittels R (Kapitel 3.7). Für die restlichen Artikel beziehen sich die Beiträge jeweils gleichermaßen auf alle Teile.

Tabelle 1. Tabellarische Auflistung der Artikel

Kapitel	Aritkel	Faktor
2	Tekles, Alexander, Katrin Auspurg, and Lutz Bornmann. 2022. "Same-gender citations do not indicate a substantial gender homophily bias." *PLoS ONE* 17(9): e0274810. DOI: https://doi.org/10.1371/journal.pone.0274810 Beiträge (CRediT): Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing Gewichtung / Anteil: 2 / 60%	1,2
3	Bittmann, Felix, Alexander Tekles, and Lutz Bornmann. 2022. "Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example." *Quantitative Science Studies* 2(4): 1246–1270. DOI: https://doi.org/10.1162/qss_a_00158 Beiträge (CRediT): Software, Data curation, Writing – review & editing Gewichtung / Anteil: 1,5 / 20%	0,3
4	Tekles, Alexander, and Lutz Bornmann. 2020. "Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches." *Quantitative Science Studies* 1(4): 1510–1528. DOI: https://doi.org/10.1162/qss_a_00081 Beiträge (CRediT): Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft Gewichtung / Anteil: 1,5 / 80%	1,2
5	Tekles, Alexander. 2022. "Gender and scientific output: How productivity, citation impact and journal prestige differ between female and male researchers". (Working Paper) Beiträge (CRediT): Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing—original draft Gewichtung / Anteil: 1 / 100%	1
	Summe	3,7

1 Summary and overview

Alexander Tekles

1.1 Introduction

Gender differences in science have constantly attracted significant attention and strong opinions in the scientific community. For example, a publication studying gender differences in fundamental physics by Strumia (2021) engendered harsh criticism (Andersen et al., 2021; Singh Chawla, 2019; Thelwall, 2021). Among other conclusions, the study suggested that biological differences between women and men play a role in gender differences in science. The study has been criticized as "methodologically flawed" (Singh Chawla, 2019), selectively citing other papers on gender differences in science and providing far-fetched conclusions based on the results (Andersen et al., 2021; Thelwall, 2021). The heated discussion about Strumia's (2021) study shows that adequate methods and careful interpretations of the results are essential when analysing gender differences in science. Moreover, prudent approaches are important because policy guidelines may be justified based on the results.

The relevance of gender differences in science is also illustrated by the persistent underrepresentation of women in science (de Kleijn et al., 2020). Several factors may contribute to it. For example, gender differences have been reported for productivity (Halevi, 2019), citations received (Larivière et al., 2013), journal prestige (Larivière & Sugimoto, 2017), collaborations (Zeng et al., 2016), mobility (de Kleijn et al., 2020), funding (Witteman et al., 2019), and chances to be hired (Moss-Racusin Corinne et al., 2012).

This dissertation's analyses focus on gender differences in citation behaviour and bibliometric indicators. Previous studies on this topic have often provided inappropriate conclusions due to insufficient awareness of relevant mechanisms and inadequate methods for considering these mechanisms in empirical analyses. This dissertation addresses these issues by introducing, evaluating, and applying methodological approaches for analysing gender differences in science.

The dissertation includes four studies. The studies in Sections 2 and 5 analyse gender homophily in citations and gender differences in bibliometric indicators measuring scientific output. The study in Section 4 compares different approaches to disambiguate author names in bibliometric data. Disambiguated bibliometric data are necessary to conduct analyses at the researcher level, as in the study in Section 5. The study in Section 3 provides an overview of matching approaches and exemplifies their use in scientometric studies. The idea of matching similar entities, on which these approaches are based, is also the basis of the methodological approaches in Sections 2 and 5.

This section provides an overview these studies and explains how they relate. Furthermore, some aspects are discussed that could not or only briefly be mentioned in the studies but are worth examining. Finally, this section summarizes the limitations of the methodological approaches in the four studies.

1.2 Theoretical perspectives on gender differences in science

Empirical analyses of gender differences in science often lack a comprehensive theoretical foundation, including (1) precisely describing the relationship that should be examined, (2) discussing theoretical arguments on how gender might relate to the variables of interest, and (3) identifying possibly relevant mechanisms and variables. However, this theoretical foundation is a prerequisite for applying adequate methods and properly interpreting the results of an analysis. Therefore, this section provides the theoretical foundation for the studies of this dissertation and serves as a basis for the methodological framework described in Section 1.3.

1.2.1 Distinguishing gender differences and gender bias

Studies concerned with gender differences in science often refer to the notion of gender bias. However, many of these studies have not explicitly and precisely stated the theoretical construct of interest, so it is unclear what they aim to measure and what conclusions their results allow regarding gender biases. In a first step, making conclusions about possible gender biases requires a conceptualization of gender biases that should be examined with the analyses.

Following the definition of bias proposed by Traag and Waltman (2022), gender bias is a direct causal effect of gender on another variable that is unjustified because it violates a normative ideal. Since this definition refers to a normative ideal, it requires assumptions about a normative frame of reference that cannot be deduced from empirical or theoretical arguments alone. A common normative framework assumed to be relevant for the science system and its actors is the ethos of science proposed by Merton (1973). In its original form, it consists of four norms:

- Universalism: The acceptance of scientific claims and the assessment of a researcher's achievements should only depend on impersonal criteria.
- Disinterestedness: Not a personal advantage but contributing to scientific progress should be the most important driver for research activities.
- Communism: Scientific achievements should be regarded as a product of and belong to the scientific community rather than single researchers who acquire recognition and esteem for their work.
- Organized scepticism: Beliefs and judgements should not be accepted without empirically and logically scrutinizing.

These norms have been used from a normative perspective to describe how researchers should behave, but they have also been used as a theoretical framework to describe the actual behaviour of researchers.

For example, the so-called normative theory of citing behaviour (Bornmann & Daniel, 2008) is based on the idea that citation decisions are governed by the norm to acknowledge the influence of other papers on a researcher's work. By regarding citations as a way to acknowledge intellectual debt (Kaplan, 1965), citation decisions can be expected to be influenced mainly through "the worth as well as the cognitive, methodological, or topical content of the cited articles"

(Baldi, 1998, p. 830). Hence, the citation decision primarily depends on the paper's content, which makes it more or less suited to be cited in a given situation. Thus, the normative theory of citing behaviour implies that citation behaviour – given the paper's content – is independent of personal characteristics like the researcher's gender.

Moreover, the theoretical perspective on the ethos of science can be used to describe which actions are regarded as deviating behaviour in the science system. This perspective can also be applied to gender differences in science, for which the norm of universalism is especially relevant. If the acceptance of a researcher's scientific achievements is only influenced by impersonal criteria, the work of female and male researchers is assessed equally. Thus, gender bias can be conceptualized as a gender difference in how a researcher's scientific work is assessed.

1.2.2 Theoretical arguments for a gender bias in the assessment of scientific work

Contrary to the assumption that researchers' behaviour closely follows the norms postulated by Merton (1973), studies that report gender differences in science may suggest that there is a gender bias in the assessment of scientific work. However, only a few theoretical arguments for such a bias can be found in the literature. A gender bias in the assessment of scientific work could be framed as taste-based discrimination concerning the researcher's gender (Becker, 1971). This would simply mean to assume that the gender bias is inherent to researchers' decisions when assessing scientific work without further explaining it.

Gender roles may provide such an explanation of a gender bias in the assessment of scientific work. Gender roles may lead to expectations about the intrinsic quality of scientific work depending on the researcher's gender. Differences in expected quality may then lead to differences in how the work of female and male researchers is assessed (Knobloch-Westerwick & Glynn, 2013). Since gender roles are difficult to observe, this mechanism is difficult to test empirically. Thus, scarce empirical evidence supports the argument that gender roles lead to a gender bias in the assessment of scientific work.

A gender difference in how research is assessed requires researchers' awareness of an author's gender. However, it is questionable whether this is always the case. For example, the authors of a paper may not be known, and it may not be possible to infer their gender based on the authors' names (e.g., only initials of the first names may be given). In this case, gender bias could not play a role in the assessment of the paper. Another argument against the existence of a gender bias is that much relevant information is available in the form of the scientific work itself when assessing it (e.g., a paper that can be assessed). Thus, a potential taste for discrimination would need to overturn the influence of the relevant information on assessing scientific work to effectively manifest a gender bias.

Even more conditions must be given for a theoretical deduction of gender homophily bias, as examined in the study in Section 2, which analyses whether researchers tend to cite researchers of the same gender. Not only is a gender difference in the assessment of scientific work necessary for such a homophily bias, but this bias would also need to differ between female and male

researchers. Hence, some form of in-group bias would need to affect the assessment of scientific work. However, the literature on gender homophily in citation decisions does not provide arguments for it. Despite the lack of theoretical arguments for a gender homophily bias in citations, many studies report results that seem to suggest such a bias. Therefore, the goal of the study in Section 2 is to test whether this finding also holds with more sophisticated methods than previous studies have used.

1.2.3 Mediating mechanisms causing gender differences

Possible biases in the assessment of scientific work can be empirically analysed by means of experiments or surveys (e.g., Bornmann et al., 2021; Bornmann, Haunschild, et al., 2022). Analyses solely based on bibliometric data cannot directly test for gender biases in the assessment of scientific work, because the data do not contain information about the assessment of scientific work itself. As an alternative, gender differences in publication- and citation-based indicators can be used because the assessment of a researcher's work manifests itself in publications and citations. Publications (especially in reputable journals) must pass a peer review process where manuscripts are evaluated, and citations are a form of acknowledgement by peers in the science system. However, it is crucial to thoroughly argue which conclusions about gender biases this approach allows and be aware of its limitations.

An important issue in this regard are mediators leading to differences in the number and impact of the papers that female and male researchers publish. These mediating mechanisms may imply biases themselves (i.e., unjustified discrimination between female and male researchers) beyond a possible bias in the assessment of the researchers' papers. For example, gender biases may lead to less funding, more teaching responsibilities, or inordinate household work for women. These differences imply fewer resources (e.g., less time or money) for female researchers to invest in research activities, which may lead to publishing fewer papers and a lower citation impact of the papers. Such biases occur before the scientific work is even assessed. Thus, differences in the work of female and male researchers can lead to gender differences in bibliometric indicators not based on gender bias in the assessment of scientific work.

An important mechanism possibly leading to gender differences in bibliometric indicators is that female and male researchers tend to be active in different disciplines (de Kleijn et al., 2020), while disciplines also differ regarding publication and citation cultures. The gender-specific segregation in different disciplines may result from two mechanisms: female and male researchers may have different chances to succeed in pursuing a career in some disciplines (Cheryan et al., 2017), or they may have different interests (Kuhn & Wolter, 2022) and therefore choose to be active in different disciplines (Ceci et al., 2014). Whatever the cause for the gender-specific segregation in disciplines may be, such a gender difference in disciplines is also supported by various empirical results (e.g., Ceci et al., 2014; de Kleijn et al., 2020; Huang et al., 2020; Larivière et al., 2013).

Differences in publication and citation cultures between disciplines lead to a different level in the number of publications per researcher or the average number of citations a paper receives. Hence, bibliometric indicators are not comparable between female and male researchers in the sense that similar indicator values can be expected, even if their work is assessed equally. For example, if female researchers predominantly publish in discipline A and male researchers predominantly publish in discipline B, while researchers tend to publish more papers and their papers receive more citations in discipline B than in discipline A, then female researchers publish fewer papers and receive fewer citations only due to the gender-specific segregation in the disciplines, even without gender bias. Therefore, to measure gender bias, only researchers or papers from similar disciplines should be compared. Otherwise, gender differences in the indicator values would be justified according to the norm of universalism and not be an indicator of gender bias.

Like the gender-specific segregation in disciplines, there may be further mediating mechanisms between gender and bibliometric indicators not caused by gender biases in the assessment of scientific work. For example, the empirical analyses in Section 5 consider academic age and publication years in this regard. To accurately measure the extent of a possible gender bias in the assessment of scientific work, all these mechanisms had to be controlled for in the empirical analyses.

However, the main focus of the studies in Sections 2 and 5 lies on gender differences in the disciplines where researchers are active. The existing empirical evidence on the gender-specific segregation into disciplines suggests that it is an important mediator (Ceci et al., 2014; de Kleijn et al., 2020; Larivière et al., 2013). Hence, the methodological approaches in Sections 2 and 5 focus on adequately controlling for disciplines. However, they are not designed to model all relevant mechanisms between gender and the bibliometric indicators.

Figure 1-1 illustrates the mechanisms leading to gender differences in bibliometric indicators. To consider all relevant mechanisms between gender and the bibliometric indicators, all variables summarized under “other mediators” had to be controlled for in the empirical analyses. Not considering all these mechanisms means a precise measurement of the degree of gender bias is impossible. However, the results in Sections 2 and 5 still allow for conclusions about the role of mediating mechanisms in general and the gender-specific segregation into disciplines in particular for analysing of gender biases in science, as well as how they can be considered methodologically.

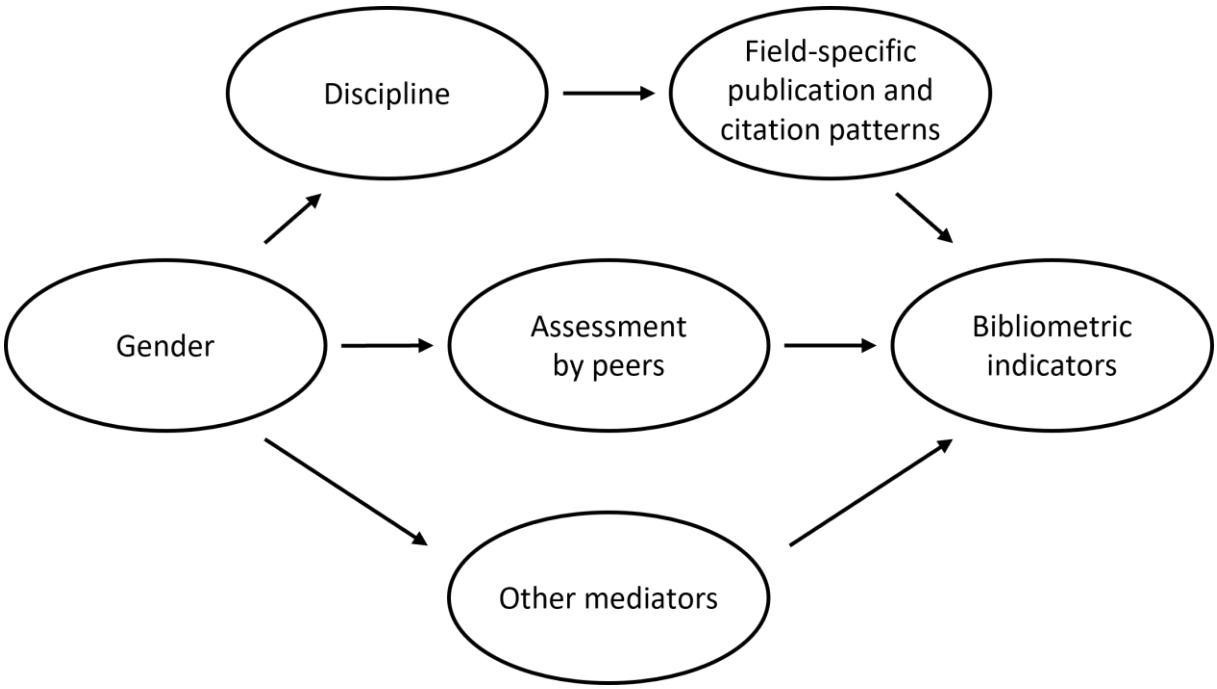


Figure 1-1. Schematic illustration of the mechanisms leading to gender differences in bibliometric indicators.
Gender bias, according to the definition in Section 1.2.1, is represented by the effect of gender on assessments by peers. Controlling for disciplines is the main focus of the studies in Sections 2 and 5. Other mediators are not differentiated further because they are not considered (or only play a minor role) in the empirical analyses.

1.3 Data and methods

This section gives an overview of the general methodological framework of the studies in this dissertation. First, Sections 1.3.1 and 1.3.2 describe the data sources used in the studies and highlight their limitations. Next, Section 1.3.3 outlines a novel approach to control for disciplines when analysing gender differences in science used in the studies in Sections 2 and 5. Finally, Section 1.3.4 describes how the gender of authors can be inferred based on bibliometric data, a prerequisite for bibliometric analyses on gender differences in science.

1.3.1 Bibliometric databases used for the empirical analyses

While all empirical analyses in this dissertation are based on bibliometric data, different data sources were used. The different data sources provide diverse information and therefore complement each other. For most studies in this dissertation, an in-house database was used, which is a pre-processed version of the Web of Science (WoS). The WoS is a standard bibliometric database containing various information about papers and their citation relations (Visser et al., 2021). This in-house database already contains several bibliometric indicators.

In addition to the WoS, the study in Section 2 is based on the Faculty Opinions database. The Faculty Opinions database is a post-publication database for papers from Biology and Medicine. The empirical analyses in Section 5 are based on Scopus, another standard bibliometric

database (Visser et al., 2021). Scopus has been chosen for this study because the author identifier included in the database was used to analyse the data at the researcher level.

An essential difference between these bibliometric databases is their coverage of publications. Both the WoS and Scopus are multidisciplinary databases, including all papers from selected journals of various disciplines. Nevertheless, both databases are biased towards "Natural Sciences and Engineering as well as Biomedical Research to the detriment of Social Sciences and Arts and Humanities" (Mongeon & Paul-Hus, 2016). The process of selecting journals to be indexed by the databases differs between the WoS and Scopus, resulting in a different number of journals indexed by the databases: roughly 21,900 journals are indexed by the WoS (Clarivate, 2022) and 25,100 by Scopus (Elsevier, 2020). The process for selecting journals to be indexed in the two databases depends on various journal characteristics, some manually assessed by reviewing the journals' content (e.g., regarding the quality of scholarly content or the readability of articles). This procedure makes it difficult to determine whether there is a difference in the characteristics of the journals indexed by the WoS and Scopus based only on the selection processes as they are described by Clarivate and Elsevier.

However, a good understanding of differences in the coverage of journals is necessary to assess how comparable results of bibliometric analyses are if they are based on the two databases. Empirical analyses on the databases coverage suggest a large overlap of the journals indexed by the WoS and Scopus. Since Scopus covers more journals than the WoS, this overlap means that most (99%, according to Singh et al., 2021) of the journals indexed by the WoS are also indexed by Scopus. However, there are slight differences regarding the disciplines covered by the databases. According to Mongeon and Paul-Hus (2016), journals from the Natural Sciences are overrepresented in the WoS compared to Scopus, whereas a larger share of the journals indexed in Scopus is from Biomedicine or the Social Sciences compared to the WoS. Paper-level analyses on the coverage of bibliometric databases show that more papers are included in Scopus than in the WoS, based on the larger number of journals indexed by Scopus (Stahlschmidt & Stephen, 2022; Visser et al., 2021). Among papers of the document types article or review, almost all papers included in the WoS are also included in Scopus, whereas more significant differences between the databases have been reported for other document types (Visser et al., 2021). Visser et al. (2021) also showed that the differences in the coverage between Scopus and the WoS can primarily be attributed to differences among papers receiving only a few citations, while both databases cover highly cited papers in most cases.

In contrast to the WoS and Scopus, the Faculty Opinions database is limited to papers from Biology and Medicine, so the Faculty Opinions database is not as comprehensive as the WoS and Scopus in terms of disciplines. However, the Faculty Opinions database contains some information on the papers manually assigned by researchers who are experts on the topics of the papers. This information was used in the empirical analyses in Section 2, which would have been impossible using only WoS (or Scopus) data.

Besides the restriction to papers from Biology and Medicine, the Faculty Opinions database only includes papers that experts have recommended (Waltman & Costas, 2014). Therefore, the papers included in the Faculty Opinions database are selective in their quality (which may be correlated with their citation impact and the prestige of the journal in which they have been published).

All these aspects must be considered in the interpretation of the empirical results. Further limitations of the datasets used in this dissertation result from the particular methodological approaches applied in the studies. For example, only researchers with at least three papers were included in the analyses in Section 5, which may have led to more senior researchers in the sample. For the analyses in Section 2, only papers for which all authors could be assigned to a gender group were considered, which may have resulted in a selective sample regarding the authors' countries of origin, which the likelihood of inferring gender depends on (see Section 1.3.4).

1.3.2 Indicators used for the empirical analyses

For the analyses on gender differences in Sections 2 and 5, different bibliometric indicators were used. The gender difference in the share of male-authored citing papers was used in Section 2 to measure the degree of gender homophily in citations. Citations generally play an important role in science because citation impact is often regarded as scientific success and relevant for scientific careers (Kamrani et al., 2020; Thelwall et al., 2020). The importance of citation impact for researchers is based on the idea that citations acknowledge intellectual debt (Kaplan, 1965), so citations indicate researchers' contributions to the scientific progress. Just like citations, the number of publications and the prestige of the journals in which researchers have published are also relevant for scientific careers, which were used in the analyses in Section 5.

While citation impact, productivity, and journal prestige play an essential role in science, other perspectives on researchers' contributions to science can also provide valuable insights. For empirical analyses, taking into account further contributions would require using different indicators than the analyses on gender differences presented in Sections 2 and 5. For example, indicators measuring the novelty of a publication (Bornmann et al., 2019; Lee et al., 2015; Uzzi et al., 2013; Wang et al., 2017) or the type of citation impact of a paper instead of simply counting how many citations a paper receives (Bornmann, Devarakonda, et al., 2020a, 2020b; Funk & Owen-Smith, 2017; Wu et al., 2019) could be used (see also Section 1.4.2).

While these alternative indicators may provide a more comprehensive picture of a researcher's contributions, they have not been established in bibliometric research and it is unclear whether they play a critical role in researchers' careers. Therefore, the empirical analyses in Sections 2 and 5 do not include these new types of indicators, instead focusing on ordinary bibliometric indicators. Nevertheless, the existence and emergence of different indicators show that the ordinary bibliometric indicators used in the study in Section 5 can only provide a limited

perspective on researchers' output. The focus on publication- and citation-based indicators also ignores other contributions to the science system, such as teaching responsibilities and funding acquisition, where gender differences may also play a role. Thus, this dissertation can only provide a limited perspective on gender differences in science.

Bibliometric analyses can refer to different levels of analysis, which implies to calculate bibliometric indicators at the corresponding aggregation level. The analyses can refer to authorships, papers, persons, institutions, or countries. Since bibliometric data are usually available at the paper level, studies using these data mainly refer to the papers as units of analysis. The empirical analyses on gender homophily in citations in Section 2 also refer to the paper level because the data only allow attributing citation decisions to author teams (which correspond to the paper level).

Another possibility would be to refer to the authorship level, which would mean operationalizing gender not for author teams, but to consider each author of a paper separately. This approach was used by Mcelhinny et al. (2003) to examine gender homophily in citations. Simply considering authorships as units of analysis would avoid ambiguities in the operationalization of gender at the paper level, which occur if not all co-authors of a paper can be assigned to the same gender. However, a citation link between two authorships would not consider the co-authors' influence on the citation decisions.

Analyses at the paper level allow conclusions about differences between female- and male-authored papers but not necessarily between female and male researchers. Regarding gender differences in scientific output, both analysis levels can be reasonable. For example, citation impact can be measured at the paper or person level. Analyses at the paper level would focus on whether female-authored papers are cited more or less often than male-authored papers. By contrast, analyses at the person level would focus on whether female or male researchers receive more citations. This perspective may be more appropriate for conclusions about the chances of succeeding in pursuing a scientific career (assuming citation impact is relevant for scientific careers). Since scientific careers refer to the person level, the citation impact must also be considered at the person level if its effect on careers is of interest.

1.3.3 Pair-based similarities to control for disciplines

The studies in Sections 2 and 5 are based on a novel methodological approach to control for disciplines in bibliometric analyses. As argued in Section 1.2, this issue is important when analysing gender differences in science. Ordinary approaches to control for disciplines in bibliometric analyses assign each entity (i.e., a paper or researcher) to one or a few disciplines based on a field classification system (e.g., Boekhout et al., 2021; Ghiasi et al., 2018; Huang et al., 2020). The disciplines can then be controlled for by only comparing the entities within a discipline. The most important field classification systems are those provided in the WoS and Scopus databases (Waltman & van Eck, 2012). These classification systems are based on journal sets,

which means that journals are assigned to disciplines based on their scope, and all papers in a journal are assigned to the discipline of this journal.

While these field classification systems are the most widely used, they have the disadvantage that many journals cover a broad range of topics (Haunschild et al., 2022; Milojević, 2020; Waltman & van Eck, 2012), also illustrated by the relatively small number of disciplines provided by the field classification systems of the WoS and Scopus. Indeed, they provide only a few hundred disciplines, whereas other field classification systems distinguish between several thousand disciplines (Waltman & van Eck, 2012). Consequently, an accurate differentiation of scientific communities with their distinct citation and publication cultures may be impossible with this approach.

Several approaches based on publication networks have been proposed as an alternative to field classification systems based on journal sets. The networks are constructed, for example, utilizing citation links or lexical similarities between the papers based on keywords, titles, or abstracts (Thijs, 2019). The idea behind these approaches is that scientific communities are represented by clusters in the networks which can be interpreted as disciplines. Besides using keywords for constructing networks, keywords can also be regarded as a (usually a rather granular) field classification system.

Such approaches may be more flexible than journal-based approaches in assigning papers to disciplines as they operate at the paper level. Nevertheless, the papers are still clustered into distinct groups. Thus, intra-group variations and inter-group similarities between the papers may not sufficiently account for the multidisciplinary character of research. Research often draws on different scientific communities, so different citation and publication cultures may be relevant for papers assigned to the same discipline. Furthermore, when analysing distinct disciplines separately, only the influence of one discipline is controlled for at a time, whereas several disciplines may be relevant for a paper.

Even if the influence of all disciplines is controlled for at once (e.g., in regression analyses with binary variables for all disciplines), the influence of each discipline can only be considered independently of other disciplines. However, research may be better characterized by certain dependent combinations of disciplines. In regression analyses, this would mean to include all interaction effects between the binary variables representing the disciplines. Due to the large number of disciplines and the curse of dimensionality, such an approach is not feasible.

Another problem with assigning papers to distinct disciplines is that it is unclear which level of granularity of the disciplines is adequate for a given dataset and research question. With a low level of granularity (i.e., a few broad disciplines), it may be impossible to control for all relevant differences between the disciplines. By contrast, a high level of granularity (i.e., many small disciplines) may lead to problems for the empirical analyses. For example, separate analyses for many disciplines would not be interpretable.

These problems of assigning papers or researchers to distinct disciplines can also be applied to the analyses in Sections 2 and 5. Papers assigned to the same discipline may be cited by groups of researchers that work on different topics, and these groups may differ regarding the gender distribution among the researchers. If this is the case, a different gender distribution among the citing papers can expected for papers assigned to the same discipline, even if no gender homophily bias exists. At the person level, a researcher's discipline must be determined based on the researcher's publications, which implies a certain degree of ambiguity because several papers must be considered, and the papers may differ in their topics. Furthermore, research interests may change over a career, making it even more difficult to assign a researcher to a particular discipline. Thus, regarding the empirical analyses on gender differences in the scientific output of researchers (Section 5), several citation and publication cultures may be relevant for a researcher and a different level of the variables of interest in these analyses (i.e., productivity, citation impact and journal prestige) can be expected for researchers assigned to the same discipline.

The methodological approach used in the analyses in Sections 2 and 5 is an alternative to using ordinary field classification systems to control for disciplines. Instead of assigning papers or researchers to distinct disciplines, this approach is based on pairwise similarities between papers regarding their topics or between researchers regarding the disciplines in which they have published (Figure 1-2). The idea behind this approach is to compare each pair of papers or researchers where the paired papers or researchers are assigned to opposite gender categories (i.e., female-authored and male-authored papers or female and male researchers are compared). For each of these pairs, the similarity between the two papers or researchers and the difference in the outcome variable (e.g., the share of male-authored citing papers or the researchers' productivity) are measured. The similarity can then be used to control for the topics of a paper or the disciplines in which researchers have published. The difference in the outcome variable can be used as an indicator for the gender differences that are analysed: the gender differences in the distribution among the citing papers' authors (i.e., gender homophily in citations) or the gender differences in output indicators.

This approach does not allow for analysing papers or researchers separately for different disciplines, as the papers and researchers are not assigned to distinct disciplines. Instead, the analyses are restricted to the most similar pairs of papers or researchers, which are the most comparable pairs with respect to the papers' topics or the disciplines in which the researchers have published. More generally, the pairwise similarities can be used for different matching approaches (see Section 1.4.2).

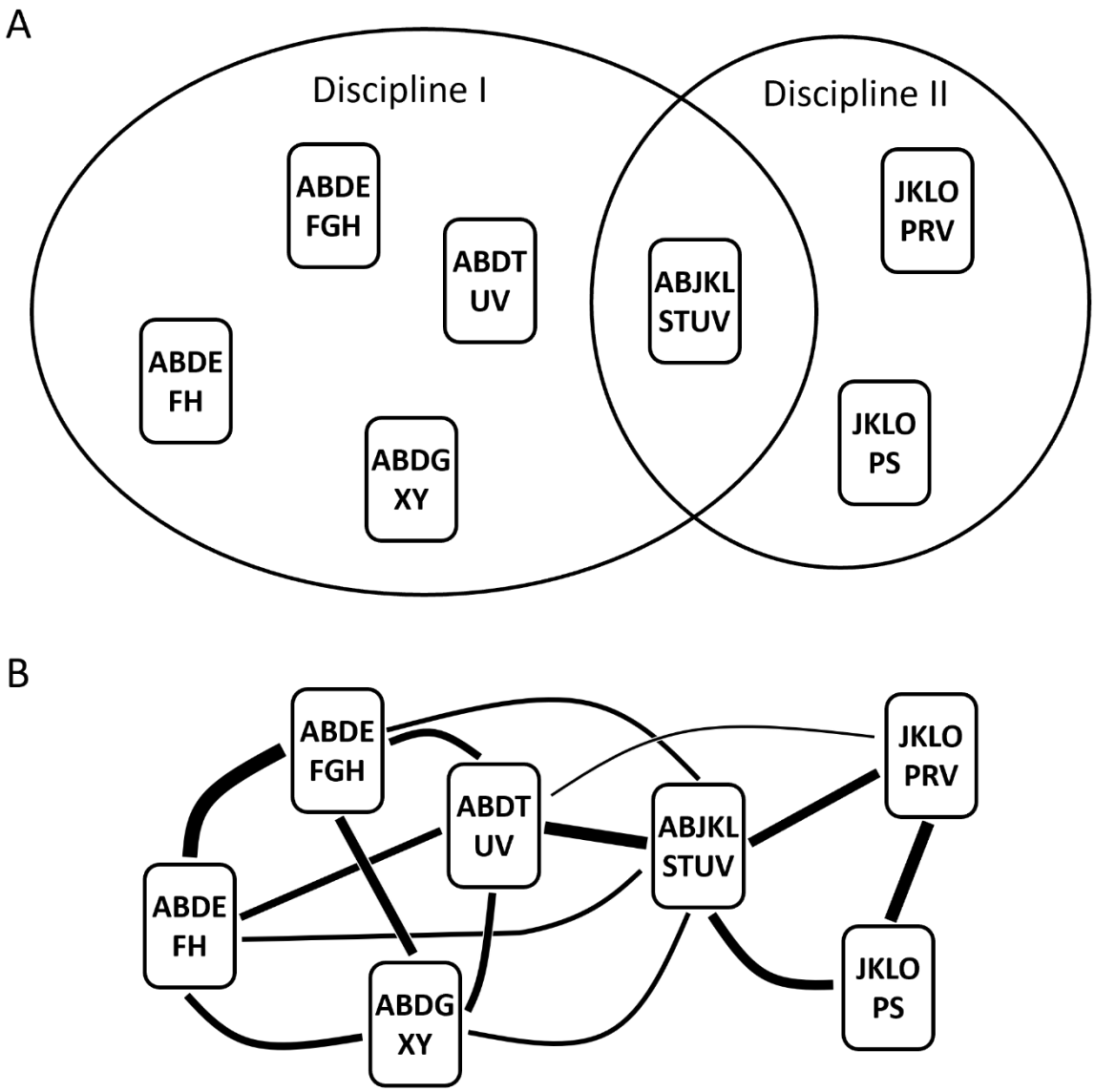


Figure 1-2. Illustration of the methodological approach using pairwise similarities instead of assigning entities to distinct disciplines.
The rectangles represent entities (i.e., papers or researchers), while the letters represent the content of a paper or the disciplines a researcher has published in. (A) Ordinary approaches assign each entity to one or a few disciplines so that similar entities are assigned to the same discipline. In the figure, two entities are assigned to the same discipline if they share at least two letters. (B) When using pairwise similarities, the similarities are calculated for each pair of papers or researchers. In the figure, the thickness of the lines represents the similarity between the entities (no lines are drawn between entities sharing no letters).

The similarity between two papers can be calculated based on citation relations or lexical similarities. Citation relations can be considered in the form of co-citations (the number of papers citing both papers of a pair) or bibliographic coupling (the number of shared cited references). A large number of papers citing both papers (i.e., co-citations) indicates that the two papers are relevant for the same scientific communities represented by the citing papers. A large number of shared cited references (i.e., bibliographic coupling) indicates that the two papers draw on similar previous work, again implying that they are relevant for similar scientific communities.

A problem when referring to co-citations for measuring the similarities is that many papers have no or only a few citations, which means that the likelihood of two papers sharing at least one citing paper is very small, even if they are similar. Due to this lack of variation, co-citations are unsuitable for controlling for disciplines. Results based on different similarity metrics in Section 2 show that the variation in the number of shared cited references is also relatively small compared to other similarity metrics.

Using the number of shared Faculty Opinions keywords (a form of lexical similarity) is the primary approach for measuring the pairwise similarities between papers for the analyses in Section 2. Due to the variation in the number of shared keywords, several levels of similarity were considered, allowing to control for the papers' topics at different levels of granularity. Another advantage of the Faculty Opinion keywords is that they are assigned by experts and can therefore be assumed to be reliable indicators for the topics of the papers (Bornmann et al., 2013).

Since the analyses in Section 5 refer to the researcher level, the similarities used in these analyses were also measured between researchers. For this purpose, the disciplines of the researchers' papers were used in the form of WoS subject categories. Thus, the aforementioned disadvantages of journal-based field classification systems indirectly apply to the similarity metric used in these analyses. However, the approach still has some advantages over the standard approach to control for disciplines at the researcher level. Instead of assigning each researcher to a particular discipline, the whole distribution of disciplines among a researcher's papers was considered to better represent multidisciplinary research activities and changes in research interests that a single discipline could not represent.

The similarity metric based on WoS subject categories can generally be replaced by other similarity metrics (for example, based on the titles and abstracts of a researcher's papers), which could mitigate the problems of journal-based field classification systems. Thus, the study in Section 5 provides a methodological approach for controlling disciplines at the researcher level that can be used with different similarity metrics. Using pairwise similarities based on the WoS subject categories is still an improvement over other approaches and therefore contributes to better understanding the gender differences in scientific output.

1.3.4 Gender inference

Bibliometric analyses of gender differences require determining the gender of authors. For small datasets, this can be achieved by manually assigning the gender to authors (e.g., Ferber & Brün, 2011; Knobloch-Westerwick & Glynn, 2013; Mitchell et al., 2013; Potthoff & Zimmermann, 2017). However, this process is impossible for large datasets like those used in this dissertation. In this case, the authors' gender must be determined automatically based on their names. Several approaches have been proposed and used in other studies for this purpose, including commercial web applications (e.g., Dion et al., 2018), open-source applications including a database to match names to a gender (e.g., Studer, 2012), national databases (e.g.,

Thelwall, 2020), and web-based data sources (e.g., Akbaritabar & Squazzoni, 2020; Ghiasi et al., 2018). A general overview of the approaches can be found in Halevi (2019).

For the empirical analyses in this dissertation, the authors' gender was determined based on the application provided by Studer (2012), which includes a database of names providing information about whether the names are typically associated with a gender in a given country of origin. This database is also used by the Python packages SexMachine (https://github.com/ferhatelmas/sexmachine/) and gender-guesser (https://github.com/lead-ratings/gender-guesser), often referenced in studies determining gender based on names. The database is published under an open-source license. By contrast, commercial web applications often rely on data not publicly available, making these applications less transparent than the database used in the application of Studer (2012).

The database also has the advantage of providing information for different countries of origin. National databases only focus on one specific country with the consequence that the gender inference based on national databases may be less reliable for other countries of origin. Since the country of origin is not available in the bibliometric data, the authors' affiliations were used instead for the analyses in Sections 2 and 5. Although the affiliation's country is not always the country of origin, this approach allows for mitigating the problem of biases regarding the country of origin, which would be impossible with a national database.

Empirical evaluations suggest that the performance of the application of Studer (2012) is comparable to other approaches (Bérubé et al., 2020; Karimi et al., 2016). A limitation of all approaches to automatically determine the authors' gender is that the gender cannot be determined for many authorships because only their initials are given in the bibliometric data. This lack of full first names especially applies to papers with early publication years, resulting in a certain selectivity regarding the publication years. A further limitation applying to all approaches is that only a binary concept of gender distinguishing between women and men can be represented in the data. However, if gender biases play a role in the assessment of scientific work, the authors whose work is assessed must be associated with a gender. If the authors are not known personally to the researchers that assess their work, the association with a gender is only possible based on the authors' names. Thus, using the authors' names to determine their gender corresponds to the research question of analysing gender biases in the assessment of scientific work.

1.4 Articles

1.4.1 Same-gender citations do not indicate substantial gender homophily bias

The first study of this dissertation examines a specific form of gender bias, a possible gender homophily bias in citation decisions. Several other studies have reported evidence for such a bias (Ghiasi et al., 2018; Potthoff & Zimmermann, 2017). The study's goal in Section 2 is to test more rigorously whether a gender homophily bias in citations exists by using more

sophisticated methods than previous studies. In particular, the papers' topics were controlled for more thoroughly than in previous studies, which have usually controlled for disciplines based on journal sets. Shedding more light on this issue seemed necessary because most studies do not provide theoretical arguments for the existence of a gender homophily bias in citation decisions. Thus, the results reported by the studies may just as well be caused by other mechanisms.

Within the theoretical framework introduced in Section 1.2, a gender homophily bias in citation decisions would mean that the gender of a paper's author (or author team) influences how other researchers assess the paper, and that this influence differs between female and male researchers who assess the paper (see Figure 1-3). This form of gender bias is in contrast to a general gender bias in citations that implies that the influence of a paper's author(s) on how the paper is assessed is the same for female and male researchers assessing the paper.

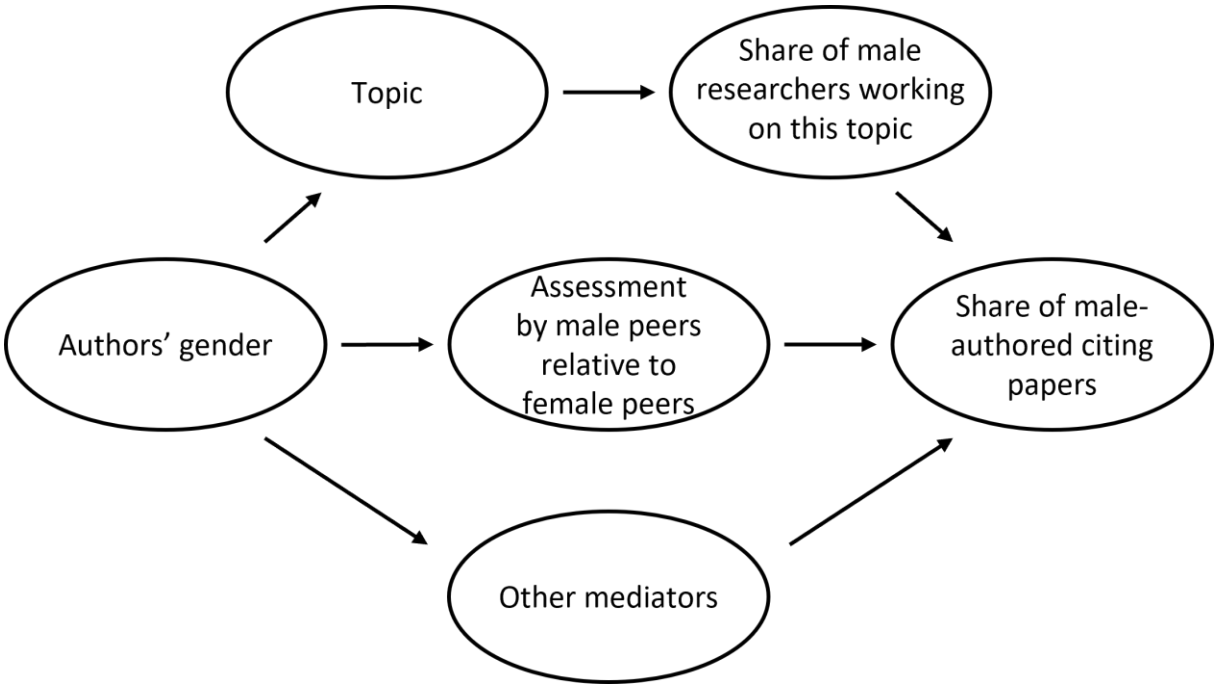


Figure 1-3. Schematic illustration of causal mechanisms leading to gender differences in the share of male-authored citing papers.

A possible gender homophily bias in the assessment of papers cannot be tested directly based on bibliometric data alone because these data do not contain information about the assessment of papers itself. Instead, the difference in the share of male-authored citing papers between male-authored and female-authored papers is used as an indicator of gender homophily in citations: a larger share of male-authored citing papers for male-authored papers than for female-authored papers suggests a homophily bias in citations. However, not only gender differences in the assessment of papers influence on the share of male-authored citing papers.

The researchers for which a paper is relevant are usually working on similar topics. Thus, the gender distribution among researchers working on similar topics as a given paper has an influence on the share of male-authored citing papers. Since male researchers are more likely to

work in domains with a large share of male researchers, it can be expected that they also have a larger share of male-authored citing papers than female researchers. Thus, without controlling for the topics, a comparison between female- and male-authored papers would inevitably show a difference in the share of male-authored citing papers, even without gender homophily bias in the assessment of the papers.

For the empirical analyses in Section 2, the papers' topics were controlled for based on the approach described earlier: pairwise similarities between the papers were calculated, and similar papers were matched. The number of shared keywords provided by the Faculty Opinions database was used for the main analyses as similarity metric. The keywords represent the topic of a paper and can therefore be used to measure the similarity of two papers' topics. Since highly similar papers are relevant for (almost) the same group of researchers, differences in the share of male-authored citing papers are not due to gender differences in the topics that researchers work on.

Besides the gender-specific selection of topics, other mediators may also influence the share of male-authored citing papers. To precisely identify the degree of gender homophily bias in citation decisions, all these mechanisms must be controlled for. One factor possibly relevant in this regard is the papers' publication years. Over the past decades, the share of female researchers has generally increased (de Kleijn et al., 2020), so male-authored papers have been, on average, published earlier than female-authored papers. It can therefore be assumed that the citing papers of male-authored papers were also published earlier than the citing papers of female-authored papers. Due to the increasing share of female researchers (i.e., a decreasing share of male researchers) over time, this means that a larger share of male-authored citing papers can be expected for male-authored papers, regardless of a possible bias in the assessment of papers.

For the empirical analyses in Section 2, publication years were controlled for to some degree. The Faculty Opinions database only includes papers published between 2002 and 2020, with almost all publication years ranging from 2006 to 2019. Thus, the variation of the publication years of the focal papers and the citing papers is limited. The limited variation of the publication years also restricts their effect on the share of male-authored citing papers because no significant changes in the gender ratio among active researchers can be expected for a short period. Other studies on gender homophily in citations have usually considered the gender distribution among the focal papers' cited references instead of the citing papers. Cited references usually span a much longer period, only limited by the coverage of the bibliometric database. Therefore, considering the citing papers is an advantage of the study in Section 2 over previous studies.

Other factors possibly relevant to the relationship between gender and the share of male-authored citing papers are the quality and the team size of the papers. The quality may depend on the authors' gender because male researchers are, on average, more senior than female researchers (Huang et al., 2020; Jadidi et al., 2018), and the paper quality may increase with the authors' seniority. Some empirical evidence also suggests that the author team size depends on the authors' gender (Ceci et al., 2014; Jadidi et al., 2018). Both the quality and the author team size

of a paper may affect the likelihood that another author will cite the paper (Beaver, 2004; Fok & Franses, 2007; van Wesel et al., 2014). This likelihood must vary with the other authors' gender to cause a difference in the share of male-authored citing papers between female- and male-authored papers (which could be misinterpreted as gender homophily bias in the assessment of the papers).

Although there are no strong arguments why the effect of a paper's quality or author team size on the likelihood that a researcher cites the paper should depend on the researcher's gender, these two factors were controlled for in an additional analysis in Section 2.5. This analysis is not based on matching similar papers but considers the papers as units of analysis. With this approach, the papers' topics cannot be controlled for as thoroughly as when matching similar papers, but it allows more control variables to be included. The papers' quality and author team size barely have an effect in this analysis, suggesting that these factors are no relevant mediators between the authors' gender and the share of male-authored citing papers. Thus, they were not considered in the analysis based on matching similar papers.

The pairwise similarity metrics used in Section 2 allows controlling for papers' topics at different levels of granularity. This flexibility allows to illustrate that the level of granularity of the papers' topics matters when analysing gender homophily in citations. The results suggest that only very similar papers should be compared to control for all gender differences in the papers' topics. An alternative would be to use more sophisticated approaches to match similar papers instead of simply restricting the analyses to the most similar pairs of papers. The study in Section 3 gives an overview of matching approaches that can be used for this purpose, and the study in Section 5 applies one of these approaches to analyse gender differences at the researcher level. All these approaches could not be implemented within the methodological framework of previous studies on gender homophily in citations as they do not consider pairwise similarities between papers.

1.4.2 Applied Usage and Performance of Statistical Matching in Bibliometrics

The study in Section 3 gives an overview of matching approaches and illustrates their use for bibliometric analyses. For this purpose, the validity of a new type of bibliometric indicators for measuring the disruptiveness of papers is examined. The analyses are based on a dataset from the journal *Physical Review E*. This dataset includes several papers of the journal that have been classified as milestone papers by the editors of the journal. The goal of the study is to compare these milestone papers with other papers from the journal not classified as milestone papers, testing whether these two groups of papers differ in the disruptiveness indicators. The idea behind this approach is that, on average, milestone papers can be assumed to be more disruptive, so the indicators should discriminate between milestone and non-milestone papers if they actually measure disruptiveness.

Several indicators have been recently proposed to expand the notion of the citation impact a paper receives. Instead of counting how many citations a paper receives, these indicators

differentiate citations to consider more information from the citation network for assessing how a paper impacts subsequent research. For example, Bu et al. (2021) proposed measuring "the depth and breadth" of a paper's citation impact. According to this conceptualization, a paper has a deep citation impact if its citing papers cite each other often and a broad citation impact if its citing papers do not or rarely cite each other (see Figure 1-4). The idea behind this conceptualization is distinguishing between papers that have an impact on many papers in one discipline and papers that have an impact on papers from several disciplines. If a paper has an impact on many papers in one discipline, it can be expected that the citing papers also cite each other (resulting in a deep citation impact). By contrast, if a paper has an impact on papers from different disciplines, it can be expected that there are fewer citation links between the citing papers (resulting in a broad citation impact).

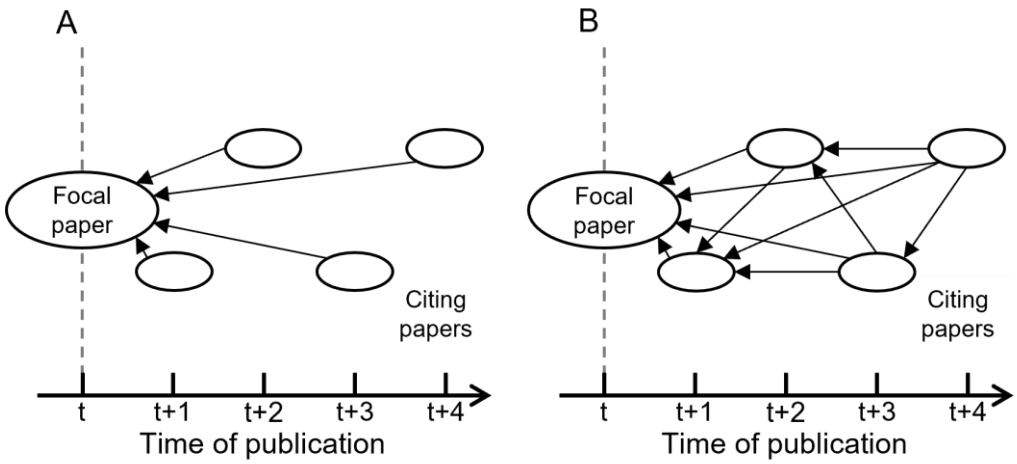


Figure 1-4. Schematic illustration of citation relations indicating broad and deep citation impact.
(A) No citation relations among the citing papers indicate a broad citation impact of the focal paper. (B) Many citation relations among the citing papers indicate a deep citation impact of the focal paper.

Bu et al. (2021) also introduced the notion of the dependency of a focal paper's citation impact, indicating whether the citing papers also refer to the focal paper's cited references (see Figure 1-5). If many citing papers refer to the focal paper's cited references, the focal paper is considered to have a dependent citation impact. In this case, the focal paper is cited alongside the papers it has cited itself. By contrast, if the citing papers do not refer to the focal paper's cited references, its citation impact on the citing papers is independent of prior work.

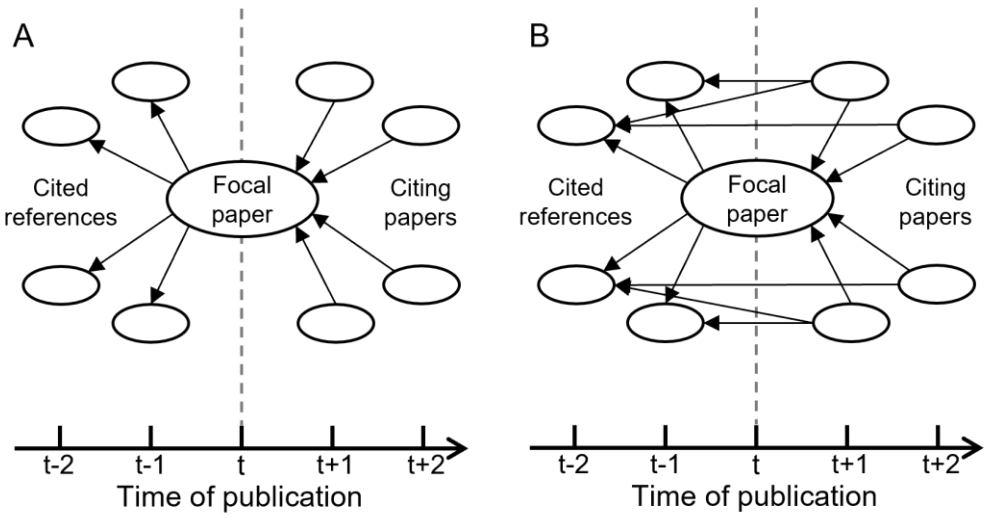


Figure 1-5. Schematic illustration of citation relations indicating the disruptiveness of a paper (dependency of its citation impact).
(A) No citation relations between the citing papers and the cited references indicate a disruptive focal paper (its citation impact is independent of prior work). (B) Many citation relations between the citing papers and the cited references indicate a consolidating focal paper (its citation impact is dependent on prior work).

This notion of the dependency of citation impact corresponds to the approach to measuring a paper's disruptiveness proposed by Funk and Owen-Smith (2017) and Wu et al. (2019).[1] Since the original version of this indicator was proposed, several studies have examined its performance and proposed modifications to overcome shortcomings in its definition (Bornmann, Devarakonda, et al., 2020a, 2020b; Bornmann & Tekles, 2019a, 2019b, 2021; Leydesdorff et al., 2021; Liang et al., 2022; Wu & Wu, 2019).

Measuring the disruptiveness of papers is not a central focus of this dissertation and was not used for the empirical analyses. However, these indicators shows that alternatives to ordinary bibliometric indicators are possible, indicating a certain demand for such alternatives. Thus, the indicators to measure disruptiveness may help to reduce the inadvertent properties of ordinary indicators. For example, the study in Section 5 suggests that ordinary indicators might differentially affect the evaluation of female and male researchers. Including new indicators (e.g., the disruptiveness indicators) could help mitigate this issue.

Besides examining the disruptiveness indicators, the primary focus of the study in Section 3 is the use of matching approaches in bibliometric analyses. So far, only a few bibliometric studies have applied matching approaches. The study intends to promote using matching approaches in scientometrics by giving an overview of several matching techniques and exemplify their

[1] Both indicators (for measuring dependency and disruptiveness) have been proposed at around the same time. They both measure a very similar dimension of citation impact but differ in terminology. Therefore, the dependency indicator proposed by Bu et al. (2021) can also be regarded as an indicator to measure the disruptiveness of a paper. The dependency indicator is also considered in the analyses in Section 3.

use. The study argues that scientometrics could profit from using matching approaches more frequently to control for variables in empirical analyses.

Matching approaches are particularly suitable for bibliometric analyses of gender differences. When analysing gender differences, two groups (female and male researchers or female- and male-authored papers) are usually compared, which is the typical setting for matching approaches. Furthermore, empirical analyses of gender differences often need to control for mediating mechanisms to avoid misinterpretations of the results, which can also be achieved by utilizing matching approaches.

Accordingly, the methodological approach in Section 2 can also be integrated into the matching framework. To match similar papers, a similarity metric between papers is necessary. Common similarity metrics for matching approaches are propensity scores based on logistic regressions or the Mahalanobis distance (King & Nielsen, 2019). For the analyses in Section 2, several other similarity metrics specifically designed to measure topical similarity between the papers were used for this purpose.

Matching approaches usually compute a counterfactual outcome for each observation so that each observation has a factual and a counterfactual outcome. Applied to the analyses in Section 2, for each female-authored paper, the share of male-authored citing papers would be calculated for the counterfactual case that the paper was male-authored. In fact, the counterfactual case cannot be observed but would be estimated based on similar male-authored papers. For male-authored papers, the counterfactual outcome would be estimated based on similar female-authored papers. The difference between the factual and the estimated counterfactual outcome could then be interpreted as the effect of the authors' gender on the share of male-authored citing papers after controlling for the similarity between the papers. This strategy is based on the assumption that if a female-authored focal paper were male-authored, it would have the same share of male-authored citing papers as similar male-authored papers and vice versa.

In contrast to standard matching approaches, no counterfactual outcomes were calculated in the analyses in Section 2. Instead, the distribution of the differences in the share of male-authored citing papers (the outcome variable) between matched papers was visualized. Hence, the differences were analysed at the level of pairs of papers instead of aggregating them at the paper level. Only for two analyses in the appendix (2.5) were the differences aggregated at the paper level. In these cases, the counterfactual outcomes were not calculated for each paper but for each paper in one of the groups (female- or male-authored papers) at a time. In the terminology of counterfactual analyses, this corresponds to the average treatment effect of the treated or the average treatment effect of the control – depending on which group of papers is considered the "treatment", and which the "control" group.

1.4.3 Author name disambiguation of bibliometric data

The study in Section 4 compares several approaches for disambiguating author names in bibliometric data aiming to match publications in the data to individual researchers. This process is

a prerequisite for bibliometric analyses at the researcher level. The challenge for author name disambiguation approaches is to solve synonyms and homonyms among author names (i.e., the occurrence of different names for the same researcher and identical names for different researchers).

Several approaches for disambiguating author names have been proposed. Some studies also provide evaluations of these approaches, but they have been performed in different settings, making it difficult to compare the approaches and judge how they perform relative to each other. Thus, the study in Section 4 aims to provide such a comparison to identify the best of the approaches included in the analyses.

Disambiguated data are necessary to analyse gender differences in science at the researcher level, as in the study in Section 5. This study uses the Scopus Author ID to identify researchers. The Scopus Author ID is an identifier provided by Scopus, which is based on an undisclosed disambiguation algorithm developed by Scopus. Some evaluations suggest that the Scopus Author ID is a reliable identifier for researchers (Aman, 2018; Baas et al., 2020; Kawashima & Tomizawa, 2015; Reijnhoudt et al., 2014), but it has not been compared with other disambiguation approaches.

Therefore, the study in Section 4 provides an evaluation framework to compare the Scopus Author ID with other disambiguation approaches. For the following analyses, the same dataset was used to compare the Scopus Author ID with the approach that produced the best results in the study in Section 4, the approach by Caron and van Eck (2014). The dataset is based on WoS data and includes only author mentions with a ResearcherID linked to the publications in the WoS. The ResearcherID is an identifier based on researcher profiles maintained by the researchers themselves.

The ResearcherID is the gold standard for researchers' publication sets in the evaluation and can be used to calculate evaluation metrics. For comparing the Scopus Author ID with the approach of Caron and van Eck (2014), the F1 metric was calculated to assess the disambiguation quality. The F1 metric ranges between 0 and 1, with larger values indicating a better quality of the disambiguated data (see Section 4.4.2 for a detailed description of the F1 metric).

Like in the analyses in Section 4, the disambiguation results were evaluated for each name block separately. A name block consists of all author mentions with the same canonical name representation of the first initial of the first name and the full surname. Separately evaluating the name blocks allows to examine how the block size affects the disambiguation quality, which is important because the disambiguation task gets more difficult with increasing name block sizes. Figure 1-6 shows the F1 values for all name blocks and both disambiguation approaches. The results suggest that the Scopus Author ID performs slightly better than the approach of Caron and van Eck (2014). Moreover, the disambiguation quality does not decline considerably for increasing name block sizes. The overall F1 values across all name blocks confirm the slight advantage of the Scopus Author ID, with an overall F1 value of 0.949 for the Scopus Author

ID and 0.900 for the approach of Caron and van Eck (2014). In sum, this comparison of the two disambiguation approaches suggests that the Scopus Author ID is a reliable author identifier, which justifies its use in the study in Section 5.

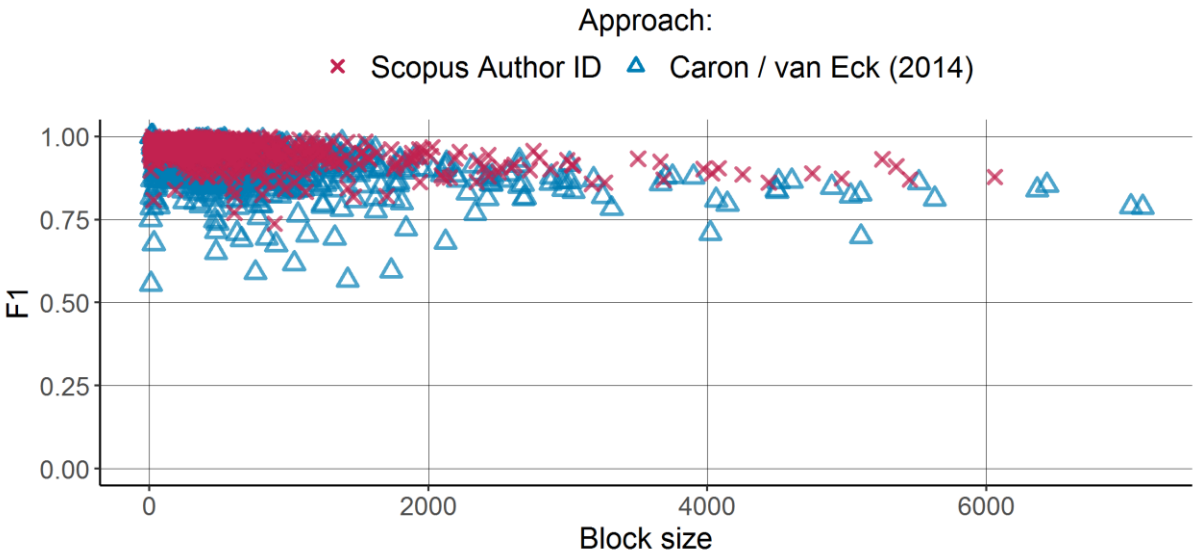


Figure 1-6. F1 values across all name blocks for the Scopus Author ID and the approach of Caron and van Eck (2014).

1.4.4 Gender differences in scientific output

The last study of this dissertation examines gender differences in bibliometric indicators measuring scientific output in terms of productivity, citation impact, and journal prestige. Whereas most of the existing studies on gender differences in bibliometric indicators refer to the paper level, the study in Section 5 refers to the researcher level. Considering different levels of analysis provides a more comprehensive picture of gender differences in scientific output by female and male researchers.

Analogous to the study on gender homophily in citations, controlling for disciplines is an essential aspect in this study. As previously mentioned, disciplines are controlled for by using a pairwise similarity metric measuring how similar two researchers are in the disciplines in which they have published (see Section 1.3.3). The similarities between researchers were used to match similar researchers. In contrast to the study on gender homophily in citations, the differences between matched researchers were not analysed at the level of pairs of papers, but the average treatment effects of the researchers' gender on the output indicators were calculated. The approach corresponds to the general procedure of kernel matching described in the study in Section 3, except a custom similarity metric was used instead of propensity scores.

Gender differences in the output indicators may result from differences in how female and male researchers' work is assessed, which would correspond to the notion of gender bias introduced in Section 1.2.1 (see Figure 1-7). However, to conclude that such a gender bias exists, all other mediators between the researchers' gender and their indicator values had to be controlled for.

Just like in the analyses on gender homophily in citations, the methodological approach in this study focuses on a few important of these mediators but is not designed to consider all relevant mechanisms.

The disciplines in which a researcher has published can be assumed to be an important factor in this regard, as discussed in Section 1.2.3. Other possible mediators controlled for in the analyses are the researchers' cohort and their academic age. Observing these variables requires researcher-level data, which is why most previous studies do not consider them in their analyses. Including the researchers' academic age also allows to test whether gender differences vary throughout scientific careers, providing a further perspective that most other studies could not examine.

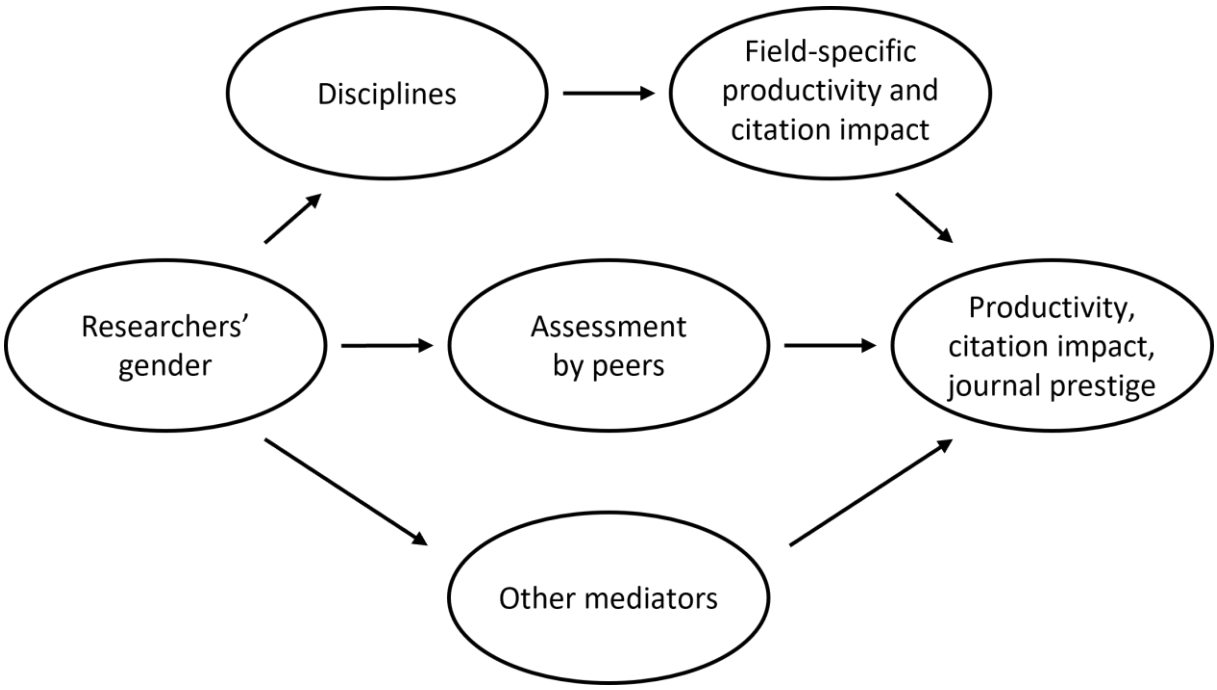


Figure 1-7. Schematic illustration of causal mechanisms leading to gender differences in output indicators.

The results show gender differences even after controlling for the disciplines in which researchers have published (including the academic age and cohort). However, the results vary for different indicators. Whereas male researchers publish more papers, female researchers achieve higher citation impact and publish in more prestigious journals. Hence, choosing a particular bibliometric indicator for evaluation tasks may lead to discrimination against female or male researchers. Assuming that the number of publications is more important for scientific careers than citation impact, as some studies suggest (e.g., Jungbauer-Gans & Gross, 2013; Kamrani et al., 2020), the differences between the indicators are more likely to result in disadvantages for female than male researchers. Furthermore, differentiating the results regarding the researchers' career length shows that many women with high potential leave the science system early in their careers.

The study in Section 5 combines the main aspects of the other studies included in this dissertation. Like the study on gender homophily in citations, it examines gender differences in science. Both studies test for gender bias in the assessment of scientific work by controlling for gender differences in disciplines as critical mediator between gender and the outcome variables. Even though the analyses cannot measure gender bias itself, they test for such a bias more rigorously than previous studies. The comparison of the Scopus Author ID with the author name disambiguation approach proposed by Caron and van Eck (2014) suggests that the Scopus Author ID allows to reliably identify researchers in bibliometric data. By using the similarities between researchers to implement a matching approach as proposed in the study in Section 3, multidisciplinary research activities can be considered better than in previous studies. Thus, the study in Section 5 generally illustrates how sophisticated methods can be used to analyse gender differences in science.

1.5 Synthesis

The studies in this dissertation provide bibliometric analyses of gender differences in science based on sophisticated methodological approaches. The results allow conclusions about gender biases in the assessment of scientific work and exemplify that a prudent theoretical foundation in combination with a sound methodological framework are the basis for accurate and meaningful conclusions.

An important finding of the analyses in Sections 2 and 5 is that the gender differences decline after controlling for disciplines. This finding supports the argumentation that controlling for disciplines is crucial when analysing gender differences in science, and that the level of granularity matters for this step. Not controlling for disciplines as thoroughly as in these analyses would therefore mean to overestimate the degree of gender bias in the assessment of scientific work. However, not all relevant mediators could be controlled for in the studies, so other mechanisms may also contribute to the gender differences found in the analyses.

Not being able to control for all relevant factors is a general problem of many bibliometric analyses because bibliometric data are observational data. Using observational data facilitates the data generation process and often allows to use large amounts of data in the analyses but implies limitations regarding the information that can be used for empirical analyses. To mitigate this limitation, as much information as possible from the data should be used to control for the relevant mechanisms in empirical analyses.

As the analyses in Sections 2 and 5 show, putting more effort in exploiting the information in bibliometric data can especially help to better consider the disciplines of papers or researchers. Bibliometric data usually contain field classifications, which have been used by other studies on gender differences in science to control for disciplines. However, whether this information suffices to represent all relevant differences in publication and citation cultures between scientific communities has not been questioned. As the results in Sections 2 and 5 suggest, the

methodological approach used in the studies of this dissertation allows to control for disciplines more thoroughly than other studies.

Since the data do not include information about how papers have been assessed and not all relevant mechanisms could be controlled for in the analyses, the results cannot be interpreted as a precise measurement of the degree of gender bias in the assessment of the papers. Nevertheless, the analyses test more rigorously than previous studies whether gender differences may be caused by a gender bias in the assessment of scientific work. The results in Section 2 show no or only a small degree of gender homophily in citations. This suggests that citation decisions are not influenced by a gender homophily bias, unless there are any mechanisms that lead to a heterophily pattern in citation decisions (i.e., researchers are more likely to cite researchers of the opposite gender) suppressing a possible gender homophily bias. A heterophily in citations could occur, for example, if more female researchers are in an early career phase than male researchers, and young researchers are more likely to cite senior researchers (e.g., because they are more proficient in the mostly male-authored established standard literature of a field than in cutting-edge research). This hypothetical mechanism shows that not all possibly relevant mechanisms were considered in the analyses. Therefore, even with the methodological approach to control for disciplines more thoroughly than previous studies, the results in Section 2 do not prove beyond doubt that citation decisions are free from a gender homophily bias.

The results in Section 5 show gender differences in the output indicators even after controlling for the disciplines in which researchers have published. Hence, the gender differences may be caused by gender biases in the assessment of scientific work. However, just like for interpretating the results on gender homophily in citations, the analyses allow no definite conclusion about the role of gender biases for the gender differences in the output indicators. Other factors like gender differences in funding, teaching responsibilities, or the amount of care work in households may also contribute to the gender differences in the output indicators.

The fact that controlling for disciplines changes the results in Sections 2 and 5 confirms that there is a gender-specific segregation into disciplines. This can be regarded as a reason for the underrepresentation of women in science. If the share of female researchers increased in a few particular disciplines (especially physical sciences, see de Kleijn et al., 2020), the overall gender ratio would be a lot more balanced. Thus, occupational preferences and research interests contribute to gender differences in science. Empirical evidence suggests that "gender differences in attitudes and expectations about math and science careers and ability become evident by kindergarten and increasingly thereafter" (Ceci et al., 2014). To better understand gender differences in science and effectively mitigate them, research should therefore not only focus on mechanisms within the science system.

The results in Section 5 show that gender differences depend on the choice of bibliometric indicators. Thus, the selection of indicators may differentially affect the evaluation of female and male researchers. Using different indicators could mitigate this issue by providing a more comprehensive picture of a researcher's output. The challenge of choosing adequate

bibliometric indicators for evaluating researchers also points to a more general problem: can bibliometric indicators even identify the "best" researchers? Bibliometrics usually rely on data that have not been generated with the intent to evaluate researchers. Hence, the dimensions of scientific output that can be measured with bibliometric indicators are arbitrary to some degree.

Therefore, the validation of bibliometric indicators is still an important task to understand what they measure and whether they allow a meaningful evaluation of researchers. Bibliometric indicators can be validated based on external criteria like the milestone assignments used in the study in Section 3, but also other study designs besides only analysing bibliometric data can help to validate bibliometric indicators. For example, simulation studies (e.g., Bornmann, Ganser, et al., 2022; Bornmann, Ganser, et al., 2020), experiments (e.g., Bornmann, Haunschild, et al., 2022; Larivière & Gingras, 2010), or surveys (e.g., Bornmann et al., 2021; Teplitskiy et al., 2022) have been used for this purpose.

Such alternative study designs could also be applied to the analysis of gender differences in science. Since it is a complex topic and various mechanisms contribute to it, different approaches are necessary to get a comprehensive picture of gender differences in science. With the studies in the following sections, this dissertation aims to provide one piece of this puzzle.

References

Akbaritabar, A., & Squazzoni, F. (2020). Gender patterns of publication in top sociological journals. *Science, Technology, & Human Values*, *46*(3), 555-576. https://doi.org/10.1177/0162243920941588

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, *117*(2), 705-720. https://doi.org/10.1007/s11192-018-2895-3

Andersen, J. P., Nielsen, M. W., & Schneider, J. W. (2021). Selective referencing and questionable evidence in Strumia's paper on "Gender issues in fundamental physics". *Quantitative Science Studies*, *2*(1), 254-262. https://doi.org/10.1162/qss_a_00119

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377-386. https://doi.org/10.1162/qss_a_00019

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytical model. *American Sociological Review*, *63*(6), 829-846. https://doi.org/https://doi.org/10.2307/2657504

Beaver, D. B. (2004). Does collaborative research have greater epistemic authority? *Scientometrics*, *60*(3), 399-408. https://doi.org/https://doi.org/10.1023/B:SCIE.0000034382.85360.cd

Becker, G. (1971). *The eonomics of discrimination*. University of Chicago Press.

Bérubé, N., Ghiasi, G., Sainte-Marie, M., & Larivière, V. (2020). Wiki-Gendersort: Automatic gender detection using first names in Wikipedia. https://doi.org/10.31235/osf.io/ezw7p

Boekhout, H., van der Weijden, I., & Waltman, L. (2021). Gender differences in scientific careers. A large-scale bibliometric analysis. In W. Glänzel, S. Heeffer, P.-S. Chi, & R. Rousseau (Eds.), *Proceedings of the 18th International Conference on Scientometrics & Informetrics* (pp. 145-156). ISSI.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45-80. https://doi.org/10.1108/00220410810844150

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020a). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259. https://doi.org/10.1162/qss_a_00068

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020b). Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). *Scientometrics*, *123*(2), 1149-1155. https://doi.org/10.1007/s11192-020-03406-8

Bornmann, L., Ganser, C., & Tekles, A. (2021). Anchoring effects in the assessment of papers: The proposal for an empirical survey of citing authors. *PLoS One*, *16*(9), e0257307. https://doi.org/10.1371/journal.pone.0257307

Bornmann, L., Ganser, C., & Tekles, A. (2022). Simulation of the h index use at university departments within the bibliometrics-based heuristics framework: Can the indicator be used to compare individual researchers? *Journal of Informetrics*, *16*(1), 101237. https://doi.org/https://doi.org/10.1016/j.joi.2021.101237

Bornmann, L., Ganser, C., Tekles, A., & Leydesdorff, L. (2020). Does the hα-index reinforce the Matthew effect in science? The introduction of agent-based simulations into

scientometrics. *Quantitative Science Studies*, *1*(1), 331-346. https://doi.org/10.1162/qss_a_00008

Bornmann, L., Haunschild, N., & Tekles, N. (2022). Are there biases in decisions to tweet on scientific papers? A plea for conducting an experimental Twitter study. *Profesional de la información*, *31*(1), e310115. https://doi.org/10.3145/epi.2022.ene.15

Bornmann, L., Marx, W., & Barth, A. (2013). The normalization of citation counts based on classification systems. *Publications*, *1*(2), 78-86. https://doi.org/10.3390/publications1020078

Bornmann, L., & Tekles, A. (2019a). Disruption index depends on length of citation window. *Profesional de la información*, *28*(2). https://doi.org/10.3145/epi.2019.mar.07

Bornmann, L., & Tekles, A. (2019b). Disruptive papers published in Scientometrics. *Scientometrics*, *120*(1), 331-336. https://doi.org/10.1007/s11192-019-03113-z

Bornmann, L., & Tekles, A. (2021). Convergent validity of several indicators measuring disruptiveness with milestone assignments to physics papers by experts. *Journal of Informetrics*, *15*(3), 101159. https://doi.org/https://doi.org/10.1016/j.joi.2021.101159

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, *13*(4), 100979. https://doi.org/https://doi.org/10.1016/j.joi.2019.100979

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155-183. https://doi.org/10.1162/qss_a_00109

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden* (pp. 79-86). Universiteit Leiden - CWTS.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75-141. https://doi.org/https://doi.org/10.1177/1529100614541236

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1-35. https://doi.org/10.1037/bul0000052

Clarivate. (2022). *Web of Science coverage details*. Retrieved 20 November 2022 from https://clarivate.libguides.com/librarianresources/coverage

de Kleijn, M., Jayabalasingham, B., Falk-Krzesinski, H. J., Collins, T., Kuiper-Hoyng, L., Cingolani, I., Zhang, J., Roberge, G., Deakin, G., Goodall, A., Whittington, K. B., Berghmans, S., Huggett, S., & Tobin, S. (2020). *The researcher journey through a gender lens: An examination of research participation, career progression and perceptions across the globe*. Retrieved 20 November 2022 from www.elsevier.com/gender-report

Dion, M. L., Sumner, J. L., & Mitchell, S. M. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, *26*(03), 312-327. https://doi.org/10.1017/pan.2018.12

Elsevier. (2020). *Scopus content coverage guide*. Retrieved 20 November 2022 from https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf

Ferber, M. A., & Brün, M. (2011). The gender gap in citations: Does it persist? *Feminist Economics*, *17*(1), 151-158. https://doi.org/10.1080/13545701.2010.541857

Fok, D., & Franses, P. H. (2007). Modeling the diffusion of scientific publications. *Journal of Econometrics*, *139*(2), 376-390, Article 1483. https://doi.org/10.1016/j.jeconom.2006.10.021

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791-817. https://doi.org/10.1287/mnsc.2015.2366

Ghiasi, G., Mongeon, P., Sugimoto, C. R., & Larivière, V. (2018). *Gender homophily in citations.* Proceedings of the International Conference on Science and Technology Indicators (STI) 2018

Halevi, G. (2019). Bibliometric studies on gender disparities in science. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 563-580). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_9

Haunschild, R., Daniels, A. D., & Bornmann, L. (2022). Scores of a specific field-normalized indicator calculated with different approaches of field-categorization: Are the scores different or similar? *Journal of Informetrics*, *16*(1), 101241. https://doi.org/https://doi.org/10.1016/j.joi.2021.101241

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609-4616. https://doi.org/10.1073/pnas.1914221117

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, *21*(03n04), 1750011. https://doi.org/10.1142/s0219525917500114

Jungbauer-Gans, M., & Gross, C. (2013). Determinants of success in university careers: Findings from the german academic labor market. *Zeitschrift für Soziologie*, *42*(1), 74-92. https://doi.org/doi:10.1515/zfsoz-2013-0106

Kamrani, P., Dorsch, I., & Stock, W. G. (2020). Publikationen, Zitationen und H-Index im Meinungsbild deutscher Universitätsprofessoren. *Beiträge zur Hochschulforschung*, *42*(3), 78-98.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, *16*(3), 179-184.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). *Inferring gender from names on the web: A comparative evaluation of gender detection methods* Conference Companion on World Wide Web, Montréal. https://doi.org/10.1145/2872518.2889385

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics*, *103*(3), 1061-1071. https://doi.org/10.1007/s11192-015-1580-z

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435-454. https://doi.org/10.1017/pan.2019.11

Knobloch-Westerwick, S., & Glynn, C. J. (2013). The Matilda effect - role congruity effects on scholarly communication: A citation analysis of Communication Research and Journal of Communication articles. *Communication Research*, *40*(1), 3-26. https://doi.org/https://doi.org/10.1177/0093650211418339

Kuhn, A., & Wolter, S. C. (2022). Things versus people: Gender differences in vocational interests and in occupational preferences. *Journal of Economic Behavior & Organization*, *203*, 210-234. https://doi.org/10.1016/j.jebo.2022.09.003

Larivière , V., & Gingras, Y. (2010). The impact factor's Matthew effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology*, *61*(2), 424-427. https://doi.org/https://doi.org/10.1002/asi.21232

Larivière, V., Ni, C., Gingras, Y., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211-213. https://doi.org/10.1038/504211a

Larivière , V., & Sugimoto, C. (2017). *The end of gender disparities in science? If only it were true...* Retrieved 20 November 2022 from https://www.cwts.nl/blog?article=n-q2z294

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684-697. https://doi.org/10.1016/j.respol.2014.10.007

Leydesdorff, L., Tekles, A., & Bornmann, L. (2021). A proposal to revise the disruption indicator. *Profesional de la información*, *30*(1). https://doi.org/10.3145/epi.2021.ene.21

Liang, G., Lou, Y., & Hou, H. (2022). Revisiting the disruptive index: Evidence from the Nobel Prize-winning articles. *Scientometrics*, *127*(10), 5721-5730. https://doi.org/10.1007/s11192-022-04499-z

Mcelhinny, B., Hols, M., Holtzkener, J., Unger, S., & Hicks, C. (2003). Gender, publication and citation in sociolinguistics and linguistic anthropology: The construction of a scholarly canon. *Language in Society*, *32*(2), 299-328. https://doi.org/https://doi.org/10.1017/S0047404503323012

Merton, R. K. (1973). The normative structure of science. In R. K. Merton (Ed.), *The Sociology of Science* (pp. 267-278). University of Chicago Press.

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183-206. https://doi.org/10.1162/qss_a_00014

Mitchell, S. M., Lange, S., & Brus, H. (2013). Gendered citation patterns in international relations journals. *International Studies Perspectives*, *14*(4), 485-492. https://doi.org/10.1111/insp.12026

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213-228. https://doi.org/10.1007/s11192-015-1765-5

Moss-Racusin Corinne, A., Dovidio John, F., Brescoll Victoria, L., Graham Mark, J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474-16479. https://doi.org/10.1073/pnas.1211286109

Potthoff, M., & Zimmermann, F. (2017). Is there a gender-based fragmentation of communication science? An investigation of the reasons for the apparent gender homophily in citations. *Scientometrics*, *112*(2), 1047-1063. https://doi.org/10.1007/s11192-017-2392-0

Reijnhoudt, L., Costas, R., Noyons, E., Börner, K., & Scharnhorst, A. (2014). 'Seed + expand': a general methodology for detecting publication oeuvres of individual researchers. *Scientometrics*, *101*(2), 1403-1417. https://doi.org/10.1007/s11192-014-1256-0

Singh Chawla, D. (2019). *In decision certain to draw fire, journal will publish heavily criticized paper on gender differences in physics*. Retrieved Jan. 4 from https://www.sciencemag.org/news/2019/11/decision-certain-draw-fire-journal-will-publish-heavily-criticized-paper-gender

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113-5142. https://doi.org/10.1007/s11192-021-03948-5

Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of Science, Scopus, and Dimensions as varying resonance chambers. *Scientometrics*, *127*(5), 2413-2431. https://doi.org/10.1007/s11192-022-04309-6

Strumia, A. (2021). Gender issues in fundamental physics: A bibliometric analysis. *Quantitative Science Studies*, *2*(1), 225-253. https://doi.org/10.1162/qss_a_00114

Studer, C. (2012). GitHub repository cstuder/genderReader. (March 20, 2019), Retrieved from https://github.com/cstuder/genderReader. https://github.com/cstuder/genderReader

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2022). How status of research papers affects the way they are read and cited. *Research Policy*, *51*(4), 104484. https://doi.org/https://doi.org/10.1016/j.respol.2022.104484

Thelwall, M. (2020). Gender differences in citation impact for 27 fields and six English-speaking countries 1996–2014. *Quantitative Science Studies*, *1*(2), 599-617. https://doi.org/10.1162/qss_a_00038

Thelwall, M. (2021). Female contributions to high-energy physics in a wider context: Commentary on an article by Strumia. *Quantitative Science Studies*, *2*(1), 275-276. https://doi.org/10.1162/qss_c_00118

Thelwall, M., Abdoli, M., Lebiedziewicz, A., & Bailey, C. (2020). Gender disparities in UK research publishing: Differences between fields, methods and topics. *El profesional de la información*, e290415. https://doi.org/10.3145/epi.2020.jul.15

Thijs, B. (2019). Science mapping and the identification of topics: Theoretical and methodological considerations. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 213-233). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_9

Traag, V., & Waltman, L. (2022). Causal foundations of bias, disparity and fairness. https://arxiv.org/pdf/2207.13665.pdf

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472. https://doi.org/10.1126/science.1240474

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: how superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601-1615. https://doi.org/10.1007/s11192-013-1154-x

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20-41. https://doi.org/10.1162/qss_a_00112

Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433-445. https://doi.org/https://doi.org/10.1002/asi.23040

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378-2392. https://doi.org/https://doi.org/10.1002/asi.22748

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416-1436. https://doi.org/https://doi.org/10.1016/j.respol.2017.06.006

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, *393*(10171), 531-540. https://doi.org/https://doi.org/10.1016/S0140-6736(18)32611-4

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*, 378-382. https://doi.org/10.1038/s41586-019-0941-9

Wu, S., & Wu, i. (2019). A confusing definition of disruption. *SocArxiv Papers*. https://osf.io/preprints/socarxiv/d3wpk/

Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., Woodruff, T. K., & Amaral, L. A. N. (2016). Differences in collaboration patterns across discipline, career stage, and gender. *PLOS Biology*, *14*(11), e1002573. https://doi.org/10.1371/journal.pbio.1002573

2 Same-gender citations do not indicate a substantial gender homophily bias

Alexander Tekles, Katrin Auspurg, Lutz Bornmann

Abstract

Can the male citation advantage (more citations for papers written by male than female scientists) be explained by gender homophily bias, i.e., the preference of scientists to cite other scientists of the same gender category? Previous studies report much evidence that this is the case. However, the observed gender homophily bias may be overestimated by overlooking structural aspects such as the gender composition of research topics in which scientists specialize. When controlling for research topics at a high level of granularity, there is only little evidence for a gender homophily bias in citation decisions. Our study points out the importance of controlling structural aspects such as gendered specialization in research topics when investigating gender bias in science.

2.1 Introduction

Gender bias is an ongoing topic in science studies. There is evidence for various forms of gender differences, as a recent review in Science suggests: "Women have fewer publications … and collaborators … and less funding … and they are penalized in hiring decisions when compared with equally qualified men. The causes of these gaps are still unclear" (Fortunato et al., 2018). At the same time, some studies report evidence against the existence of gender differences, e.g. with regard to funding (Forscher et al., 2019; Ginther et al., 2016; Marsh et al., 2009) or hiring decisions (Stewart-Williams & Halsey, 2021; Williams & Ceci, 2015). One question that has been frequently investigated hitherto is whether female scientists are cited less often by male scientists than by their female peers. The existence of such gender bias would imply disadvantages for female scientists. Citation scores are increasingly applied as a core metric to evaluate the performance of individual scientists as well as the quality of faculties, departments and institutional excellence at a global level (Hicks et al., 2015). Citations also matter for the distribution of resources, such as research grants or tenured positions (Wildgaard, 2019). To achieve gender equality in science, it is thus important to monitor possible gender gaps in citations and to understand their underlying reasons.

To date, literature shows mixed evidence of a possible gender citation gap. Some studies find no gender differences in citations or that female authors receive more citations than male authors (Ceci et al., 2014; Halevi, 2019; Lynn et al., 2019; Thelwall, 2018), while other studies report that male authors receive more citations than female authors (Chatterjee & Werner, 2021; Larivière et al., 2013; Zhang et al., 2021). Be that as it may, gender homophily in citation decisions has been suggested as a reason for a possible gender citation gap where male authors receive more citations than female authors (Dion et al., 2018; Maliniak et al., 2013). We conceptualize gender homophily in citation decisions as the preference of scientists to cite other scientists only because they belong to the same gender category. We look at authors' gender expression through names, but cannot distinguish this from authors' gender identity which might differ. We also applied a binary concept of gender that only distinguishes between women and men. Thus, our analyses rely on a simplified concept of gender. Nevertheless, our analyses should provide a first insightful analysis of the extent to which preferences versus structural aspects lead to citation inequalities.

Our notion of homophily captures preferences that go beyond structural reasons for gendered citation patterns. Such structural aspects exist for example with the gendered specialization on research topics, which can result in male scientists citing more male-authored papers than female scientists (and vice versa). However, we conceptualize homophily as the "bias" that leads same gender peers to cite each other more often than what a baseline model of gender-blind selection of relevant literature would predict (McPherson et al., 2001). Evidence suggesting gender homophily in citation decisions has been reported for the fields economics (Ferber, 1986; Ferber & Brün, 2011), anthropology (Lutz, 1990), sociology (Davenport & Snyder, 1995), library and information science (Håkanson, 2005), communication science (Knobloch-

Westerwick & Glynn, 2013; Potthoff & Zimmermann, 2017), political science (Mitchell et al., 2013), and across different fields (Dion et al., 2018; Ferber, 1988; Ghiasi et al., 2018; Mcelhinny et al., 2003). See 2.5.2 for details of these studies. Given the fact that more scientists are male than female, homophily in citation decisions alone could account for the observed citation advantage for male authors: as long as men are overrepresented in science, citing along gender lines would boost up citation scores of male scientists simply for the fact that they belong to the dominant gender group (Maliniak et al., 2013).

In our study, we tested the hypothesis that gendered citation patterns can emerge on the macro level due to structural aspects alone, with no gender homophily being at play. With gendered citation patterns, we mean the fact male scientists cite male-authored papers more often than female scientists (and vice versa) when looking at the pool of all scientists, regardless of their research area. Scientists' gender is strongly related to the topic they are working on (Larivière et al., 2013). It follows that gendered citation patterns may result from varying gender distributions across different topics: whenever papers are pooled from discrete subfields that vary in their gender ratio, but which do not have one joint risk pool of papers to be cited for substantive reasons (e.g. due to their topic relevance), there will be a difference in the gender distribution among the cited authors between female and male scientists. Failure to control for the research topic as an important mediator between the gender of authors and gender distribution of cited references would then lead to an overestimation of homophily (Figure 2-1) (Holman & Morandin, 2019).

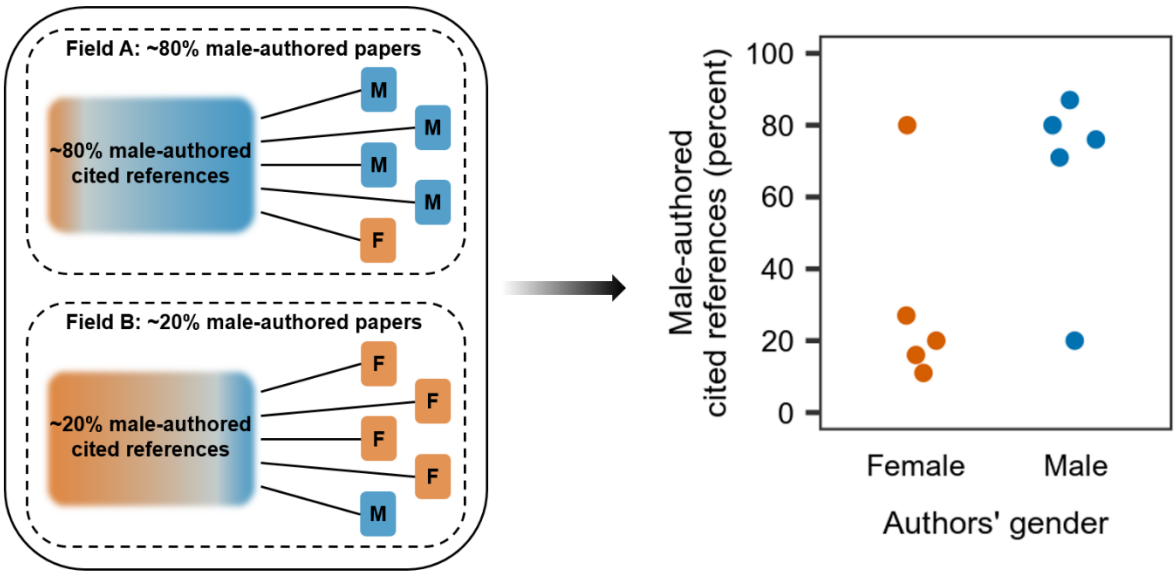


Figure 2-1. Schematic example illustrating the emergence of gendered citation patterns due to varying gender distributions across topics.
On the left, female-authored and male-authored papers (denoted with "F" and "M" respectively) in two different fields and the gender distribution among their cited references are illustrated. Papers from one field are assumed to have one joint risk pool of papers to be cited. The plot on the right shows the resulting shares of male-authored cited references on the aggregated level. Even though this difference is solely based on the varying gender distribution across topics, it may be erroneously interpreted as a gender homophily bias in the authors' citation decisions.

To identify homophily bias, it is therefore important to control for structural aspects that make some papers to more adequate sources to be cited than other papers. Besides research quality, the most important structural aspect to define such risk pools of papers is certainly topic similarity (overlap in research questions, theories and/or methods). Previous homophily studies have already tried to control for this topic similarity by considering the journals authors have published in (Davenport & Snyder, 1995; Dion et al., 2018; Ferber, 1986, 1988; Ferber & Brün, 2011; Håkanson, 2005; Knobloch-Westerwick & Glynn, 2013; Lutz, 1990; Mcelhinny et al., 2003; Mitchell et al., 2013; Potthoff & Zimmermann, 2017). However, it is questionable whether this sufficiently controls for topic similarity, because journals often accept work from different subjects that show little or even no overlap in research topics or methods. Ghiasi et al. (2018) used more information on the papers' content to identify topic similarity by matching papers that appeared in the same issue of the same journal based on the papers' abstract and title. But matching only within journal issues may have restricted the ability to identify papers that are similar due to the small number of papers: it is highly probable that more similar papers are available beyond the journal issue.

Our goal was therefore to test more rigorously whether gendered citation patterns are caused by a gender homophily bias. To precisely control for the risk pool of papers to be cited, we measured the topic similarity between papers based on (the combination of) keywords that were assigned manually by experts. We drew on data provided by Faculty Opinions (https://facultyopinions.com/; previously F1000Prime) that contains this information for papers published in 2002-2020. The keywords assigned to these papers were curated by an editorial team at Faculty Opinions in cooperation with leading scientists and clinicians in the corresponding biological and medical research fields. For each of the papers, the data also include information from reviews of experts in these fields (Bornmann, 2015). We were able to classify the gender of all authors (using a binary coding distinguishing typical male and female first names) of ~38,500 papers in that database along with ~335,000 papers that subsequently cited these focal papers. We validated the results with Web of Science (WoS) data (including almost 400,000 papers across all scientific disciplines) and alternative approaches to measure topic similarity. By controlling the topic similarity between papers at different levels of granularity, we were able to study how empirical results on gender homophily are influenced by the approach to control the papers' similarity.

Our main finding is that thoroughly controlling for research topic is important for validly assessing the degree of gender homophily. The level of observed gender homophily substantially decreases, the more fine-grained measurements of topic similarity are used. At a high level of granularity, only very little evidence remains for a possible homophily bias. We conclude that although gender homophily may affect citation decisions to some degree, the impact of this bias has likely been overestimated in the literature due to insufficient controls for topic similarity to define potential pools of papers to be cited in different research areas.

2.2 Results

2.2.1 Results on biomedicine with Faculty Opinions data

Figure 2-2 shows the results of a linear OLS regression using the papers included in the Faculty Opinions database and their metadata as observations. Table 2-1 shows the coefficient estimates for the regression analyses. The dependent variable is the share of male-authored papers among the citing papers. Note that we used the focal papers' citing papers instead of their cited references, as other studies have done. This allowed for a better control of the publication year of the papers on the cited side (the focal papers in our case). This is necessary to control for the gender composition of authors: the gender distribution in science has changed over time, which, if not taken into account, could also artificially lead to evidence of homophily bias when male authors are more likely to work in fields whose literature appeared earlier (see 2.5.2). The main independent variable is the gender of the focal papers' authors, whose direct effect can be interpreted (once all indirect effects arising from structural aspects are controlled) as the degree of gender homophily in citation decisions: if there was a gender homophily bias in citations, the share of male-authored papers among the citing papers would be higher for male-authored focal papers than for female-authored focal papers. To facilitate a clear interpretation of the results, we focused on the comparison between female-only and male-only author teams in our analyses and included other papers as "mixed-authored." We excluded all self-citations (i.e., citations where citing and focal paper share at least one author name), since they artificially increase the correlation between the gender of the focal and citing papers' authors.

Model M1 (green) shows that for male-authored focal papers, the share of male-authored citing papers is about 12.6 percentage points higher than for female-authored focal papers. However, no further variables are included in this model. Model M2 (blue) shows that this effect reduces to about 7.1 percentage points when controlling for keywords (in the form of binary variables for all keywords in the Faculty Opinions database). This means that a gender-specific selection of scientists into different topics is – at least partly – responsible for the observed gendered citation patterns. Controlling for further factors (average quality rating, age of paper, and number of authors) in Model M3 scarcely changes the effect of the gender of the focal papers' authors. We controlled for publication year and number of authors, since empirical analyses suggest that the share of female authors increased over time (West et al., 2013), and that female authors have fewer co-authors than men (Ceci et al., 2014; Jadidi et al., 2018).

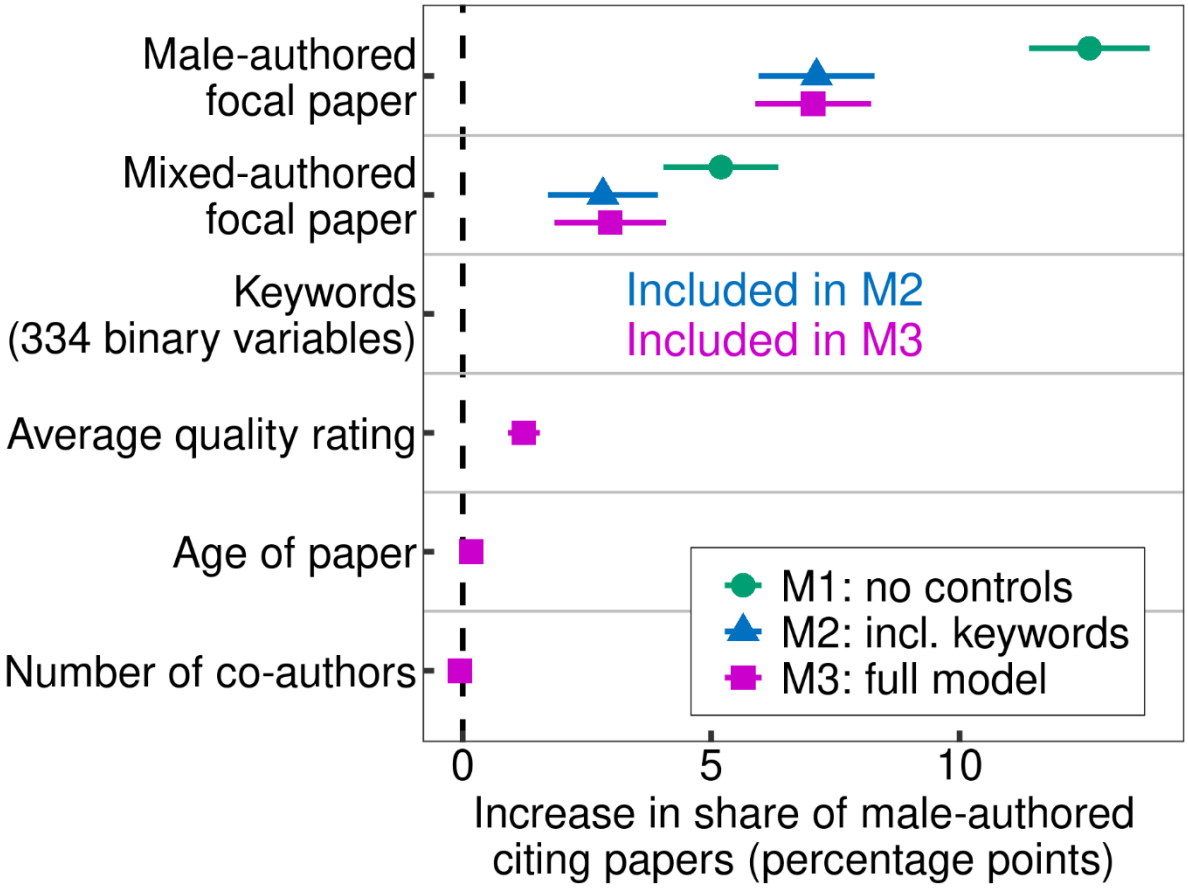


Figure 2-2. Marginal effects (with 95% confidence intervals) of three regression models on the level of focal papers.
For each model, the dependent variable is the share of male-authored papers among the citing papers. In addition to the gender of the focal papers' authors, we successively included as possible structural factors the focal papers' keywords for research topics (in the form of binary variables to control for all 334 keywords in the Faculty Opinions database), the quality rating (average quality rating in case of quality ratings by multiple experts for one paper), the age of the papers (publication year), and the number of authors. All models are based on 38,439 observations (focal papers). For more information and detailed analyses, see 2.5.1.

In line with some previous studies (Dion et al., 2018; Ferber, 1988; Ghiasi et al., 2018), these results suggest that controlling for topics is necessary in order to not overestimate the degree of gender homophily preferences in citations. The results also reveal that a certain degree of homophily remains even after controlling for topic.

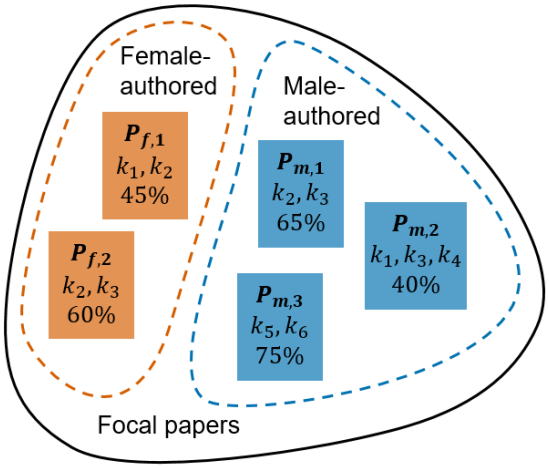
However, the inclusion of keywords in the form of binary variables in a regression model only allows controlling for each keyword independently of other keywords. Since research is usually reflected by more than one keyword (on average, 11 keywords are assigned to a paper in the Faculty Opinions dataset), topics may be better represented by certain (dependent) combinations of keywords. To consider this, we generated pairs of focal papers such that one paper is authored only by male scientists and the other paper is authored only by female scientists (see Figure 2-3). For each pair, we used the number of shared keywords as a measure for the similarity between the two papers. The difference in the share of male-authored papers among the citing papers that remains after controlling for topic similarity (measured on different levels of

granularity) serves as an indicator for gender homophily. Using these differences, we plotted histograms for all pairs of focal papers with at least *X* shared keywords. With increasing X, the pairs are increasingly similar in terms of keywords (describing both focal papers' research topic).

Table 2-1. Results for the regression models on the level of focal papers

	Dependent variable: share of male-authored citing papers		
	M1	M2	M3
Gender of focal papers' authors (reference category: female)			
Male	12.613***	7.123***	7.053***
	(0.620)	(0.596)	(0.596)
Mixed	5.198***	2.823***	2.969***
	(0.590)	(0.565)	(0.573)
Faculty Opinions keywords		Included	Included
Quality rating (average)			1.232***
			(0.164)
Age of paper			0.175***
			(0.034)
Number of authors			-0.059*
			(0.026)
Intercept	23.115***	26.281***	23.581***
	(0.579)	(0.588)	(0.695)
N	38,439	38,439	38,439
R^2	0.029	0.136	0.138

Note. Regression estimates underlying Figure 2-2. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed tests).



Pair	D	S
$P_{m,1} - P_{f,1}$	20	1
$P_{m,1} - P_{f,2}$	5	2
$P_{m,2} - P_{f,1}$	-5	1
$P_{m,2} - P_{f,2}$	-20	1
$P_{m,3} - P_{f,1}$	30	0
$P_{m,3} - P_{f,2}$	15	0

D: Difference in share of male-authored citing papers (percentage points)
S: Similarity (number of shared keywords)

Figure 2-3. Generating pairs of focal papers.
On the left, five focal papers are illustrated, together with their keywords (k_j) and the share of male-authored citing papers (in %). The table on the right shows all pairs of focal papers such that one paper of a pair is female-authored and the other paper is male-authored. Column D shows the difference in the share of male-authored citing papers for a pair, which is used as an indicator for the degree of gender homophily. Column S shows the number of shared keywords, which is used as an indicator for the similarity between two papers.

Figure 2-4 shows that the average difference in the share of male-authored citing papers between male-authored and female-authored focal papers is positive, meaning that male-authored focal papers are more likely to receive their citations by male authors than female-authored focal papers. But for increasing X (i.e., topic similarity), the difference approaches the shape of a normal distribution. The shape of a normal distribution could be expected if there is no gender homophily in citations: with no homophily bias, on average, the difference in the share of male-authored citing papers would be zero, and the differences would be distributed symmetrically around this average (with cases becoming the less frequent, the larger the distance to this zero-difference reference line). These results suggest that after controlling for the topic on a sufficiently high level of granularity (i.e., beyond the inclusion of keywords in the form of binary variables), gender homophily can be scarcely observed.

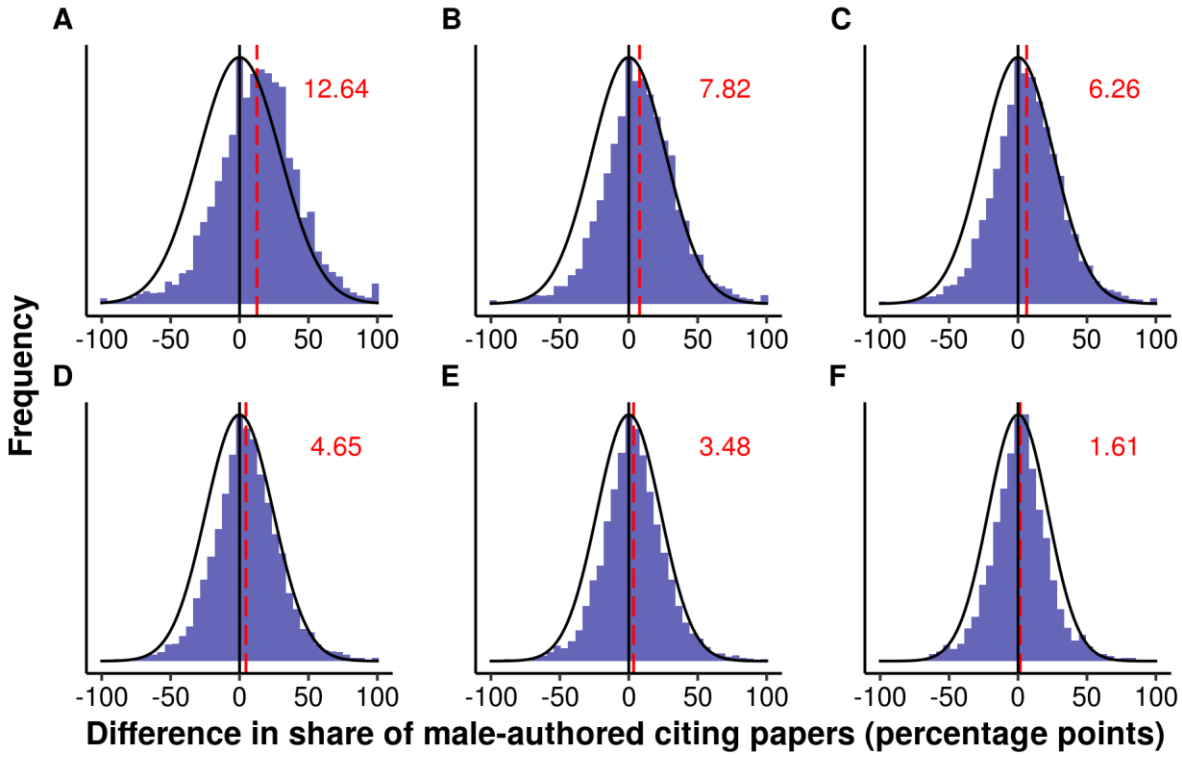


Figure 2-4. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citations than the female-authored paper of this pair. The histograms differ in terms of the minimum number of shared keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,702,080 pairs in (A), 765,642 pairs with at least one shared keyword in (B), 223,837 pairs with at least two shared keywords in (C), 58,465 pairs with at least three shared keywords in (D), 14,167 pairs with at least four shared keywords in (E), and 3,010 pairs with at least five shared keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

2.2.2 Extension to other research fields and data

We deem the keywords in the Faculty Opinions database a reliable approach for measuring the topic similarity between papers, since these keywords are based on expert knowledge and provide a more consistent measurement compared to keywords idiosyncratically chosen by authors (Bornmann et al., 2013). Although this information is a particular advantage of our dataset, the dataset is restricted to biological and medical areas and research of outstanding quality (Waltman & Costas, 2014).

We therefore tested whether our results also hold for alternative similarity measurements (Figure 2-5A-E) and a set of focal papers covering a broader range of fields than the Faculty Opinions dataset (Figure 2-5F). For each similarity measurement, we defined six similarity levels, according to the number of shared Faculty Opinions keywords used in the results shown in Figure 2-4. All of these analyses confirm the main result: the degree of observed gender

homophily decreases as the similarity between papers is controlled for more thoroughly. However, the remaining gender effects indicating homophily are generally slightly larger than when using the keywords provided in the Faculty Opinions database (see also 2.5.2). The most plausible explanation for this result is that the alternative approaches for measuring the similarity between papers provide less precise measures of topic similarity than the more standardized assignment of keywords by experts. This insufficient control for citation pools may induce spurious evidence of gender homophily.

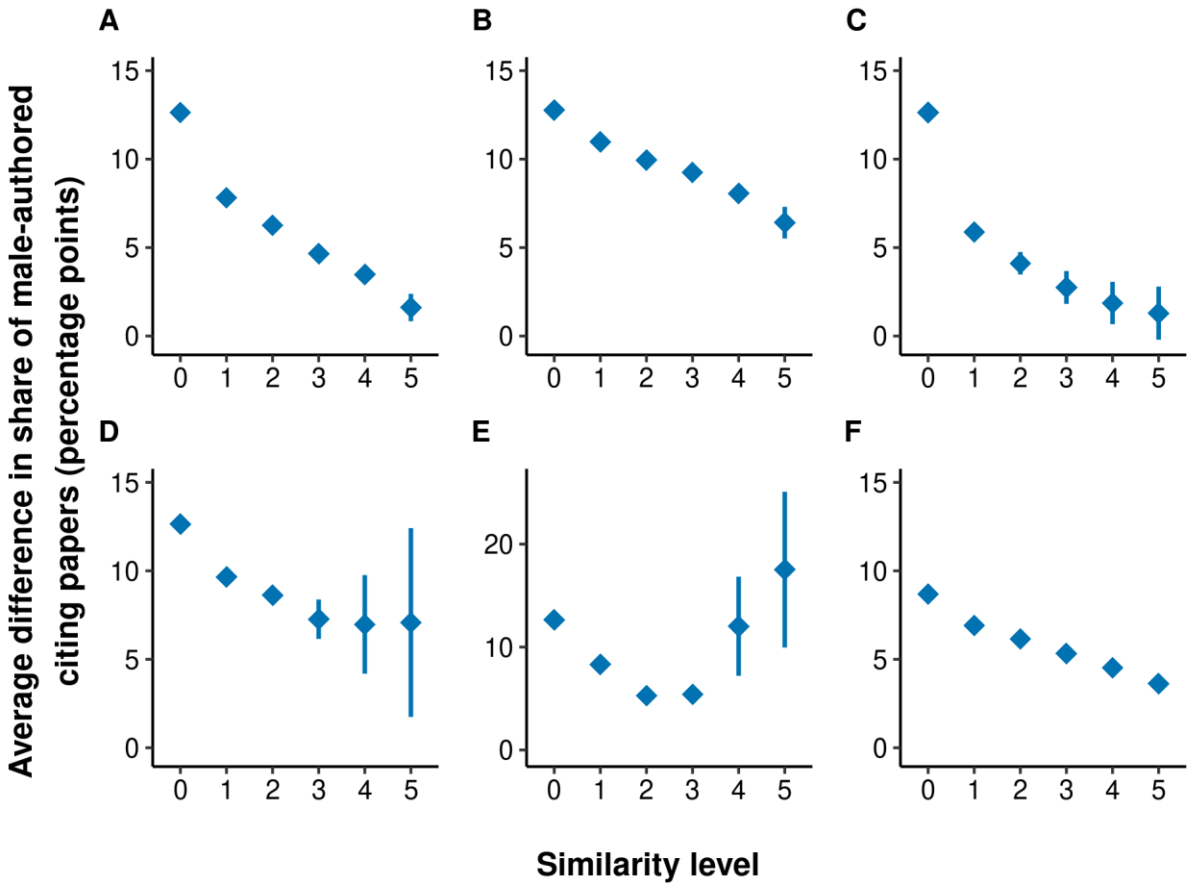


Figure 2-5. Results for alternative approaches to measure similarity.
Average differences in the share of male-authored citing papers across different similarity levels. Bars show 95% confidence intervals. (A)-(E) are based on Faculty Opinions data, (F) on WoS data. The similarity between focal papers is measured using the number of shared keywords provided by the Faculty Opinions database in (A), abstracts and titles in (B) and (F), cited references in (C), keywords provided by the WoS in (D), and WoS subject categories in (E). For (C)-(E), the similarity levels represent the number of shared keywords, cited references or subject categories. For (B) and (F), the similarity levels are specified such that the share of pairs of focal papers corresponds to these shares in (A), see 2.5.2 for detailed results.

For identifying comparable pairs of papers, the approach based on titles and abstract is the most efficient alternative to using Faculty Opinions keywords (this also manifests in the relatively large confidence intervals for the other approaches). Therefore, we used this approach to expand the analyses to all papers from the WoS published in 2012 (the mean publication year for the papers in the Faculty Opinions dataset). For these data, the observed degree of gender homophily is generally smaller than for the Faculty Opinions data. Since these results are based on the

less precise approach to measure topic similarity based on titles and abstracts, we can be even more sure that there is no marked gender homophily that goes beyond gender compositions of research fields.

We were able to replicate our results in various further robustness checks (e.g., including more controls, using alternative statistical models or using female- instead of male-authored citing papers to measure gender homophily; see 2.5.2). A noteworthy side result of these checks is that both, not excluding self-citations and using cited references instead of citing papers for measuring the degree of homophily (which means a limited control of papers' age), inflate the observed gender homophily. This suggests that the gender homophily reported in the literature is also inflated by those design aspects: only some existing studies excluded self-citations and none used citing papers to measure homophily.

2.3 Discussion

Just as in previous studies, we were unable to conduct a randomized experiment and instead had to rely on large-scale bibliometric data. The main takeaway from our study is the necessity of using adequate measures for controlling mediating factors when studying gender bias: only when all relevant mediators are controlled with exact measurements can the genuine gender bias that defines homophily be identified. Our study reveals the importance of one mediating factor in particular: the research topic. Without controlling the topic at a fine-grained level, this study would have erroneously concluded (as did others) that there is a strong homophily bias. The very small evidence for homophily bias that remains in our study after controlling for topic similarity suggests that other mediators are not very meaningful. Since previous studies have shown that more productive, senior authors collect more citations (van den Besselaar & Sandström, 2017), seniority might be a possible meaningful mediator for homophily. Based on our results, however, seniority can be excluded as meaningful mediator.

Similar to previous studies on gender homophily in citations, our approach to identify the authors' gender based on their first names implies an imprecise concept of gender. It is unclear to what extent this approach measures only a person's gender expression and not their possibly different gender identity. Thus, our results do not allow to differentiate between these notions of gender. Our approach to infer the scientists' gender is also limited to a binary concept of gender, which means that we cannot draw any conclusion about scientists with non-binary gender. Future research could address these issues by applying a more differentiated concept of gender.

Gender (homophily) bias is also suspected in many other realms of science, including reviews of publications, grant assignments or decisions to select co-authors or peers for acknowledgements (Araújo et al., 2017; Holman & Morandin, 2019; Paul-Hus et al., 2020). Reliable measures of research topics (and other possible sources of gender heterogeneity) are needed not only to rule out mediators in these realms as well, but also to achieve sufficient statistical power to detect genuine gender bias that may still exist in many realms (decisions) in science (Roper,

2019). Developing measurements of research field-specific clustering is therefore an important topic (in bibliometrics) for investigating gender bias. So far, there is no robust and generally accepted standard solution (Waltman & van Eck, 2019). Our results suggest combinations of keywords assigned by experts to be a promising approach, at least to measuring risk pools that underlie citation decisions.

Our study also points out, in accordance with many other empirical studies (e.g., Boekhout et al., 2021; Ceci & Williams, 2011; Duch et al., 2012; Huang et al., 2020), that there are structural mechanisms other than gender homophily leading to gender differences in citations. Also many other studies found no evidence for a genuine gender bias in science once they controlled for structural factors, such as different career lengths or qualifications (de Kleijn et al., 2020; Forscher et al., 2019; Ginther et al., 2016; Huang et al., 2020; Lynn et al., 2019; Williams & Ceci, 2015). In a recent blog post, Traag and Waltman (2020 Dec 10) emphasize the importance in gender bias studies of understanding the underlying causal mechanisms. Only by uncovering the micro-mechanisms actually producing the gender differences observed on the macro level can effective measures be proposed to mitigate them. Our results indicate that the sorting of female and male scientists into different fields and topics (which has been shown, for example, by Holman et al., 2018; Thelwall et al., 2020; West et al., 2013) is one of the most important mechanisms producing gendered citation patterns on the macro level. Therefore, one should in particular research the mechanisms underlying gendered specializations in research topics (so-called “horizontal segregation”), whether due to self-selection or sorting by gatekeepers.

2.4 Materials and methods

The Faculty Opinions data that we used in this study includes expert ratings of the papers’ scientific quality, which are given in the form of "good," "very good," and "excellent." Thus, only papers at a high quality level were selected for inclusion in the database. Information about the topic of the papers is given in the form of keywords assigned by experts (an editorial team at Faculty Opinions in cooperation with leading scientists and clinicians in the corresponding biological and medical research fields). There are 334 different keywords occurring in the database, and an expert may have assigned multiple of these keywords to a paper. Since keywords and quality ratings have been assigned by experts in the field (and in many cases by more than one expert per paper), we can assume a high accuracy of the data.

We matched the papers in the Faculty Opinions database (focal papers) with metadata on authors and topics from the WoS. From these data, we used the author names (to infer the authors’ gender) and the publication year for both the focal papers and all of their citing papers. For this purpose, we used an open source application for assigning a gender category (female or male) to first names (Studer, 2012; see also 2.5.1). At the paper level (for both focal and their citing papers), we operationalized the authors' gender in the form of three categories: all co-authors are female, all co-authors are male, or the team consists of both female and male co-authors. In the regression analyses, we included the gender of the focal papers’ authors in the form of two

dummy variables for the categories indicating male-authored focal papers and mixed-authored focal papers, with female-authored focal papers as reference category.

2.5 Appendix

2.5.1 Materials and methods

Dataset

For the analyses presented in Figure 2-2, Figure 2-4, and Figure 2-5A-E, we used papers that are included in the Faculty Opinions database and for which we could merge the necessary metadata from the Web of Science (WoS) database. The Faculty Opinions database initially contains information for 162,071 papers. Due to missing metadata and restrictions in our analyses, we could not include all these papers in our analyses. Figure 2-6 illustrates how many papers had to be excluded at each step in the data preprocessing phase. Most papers had to be excluded due to insufficient information on the authors' gender (for at least one author of a paper, the gender could not be inferred). We decided to use only reliable gender information, although this reduced the number of papers that could be included in the analyses. For a large share of authors, the gender could not be inferred because no first name or only the initials of the authors' first name(s) are given. Of all authors with missing gender information, 1.9% have no author name provided in the WoS, 28.6% have only one letter given as first name string, and another 19.3% have only two letters given as first name string. This suggests that for at least one-third of the authors the missing gender information is due to a missing full first name: for almost all cases with only one letter (and for most cases with two letters) given, it can be assumed that these are the first names' initials.

On the paper level, we labelled the gender as missing if at least one author has missing gender information. Among the papers with missing gender information, 36.7% have at least one author with missing name and 73.5% at least one author with unisex name (see next section for more information on how the authors' gender was classified).

Further analyses show that information on gender is missing for papers with older publication dates in particular. Papers with sufficient gender information have on average been published in 2012, while papers with insufficient gender information have on average been published in 2011. This can be expected due to the increasing availability of full first names in the WoS over time. Since even one author with missing gender information results in no gender being assigned to a paper, papers with multiple authors are more likely to have missing gender information. This is also reflected in our data: papers that could be assigned a gender category were written on average by fewer authors (arithmetic mean: 6; median: 5) than papers that could not be assigned a gender category (arithmetic mean: 10; median: 7). Papers without gender information also have more citations on average (arithmetic mean: 126; median: 57) compared to papers with gender information (arithmetic mean: 92; median: 43). However, this difference reduces and reverses (papers without gender information receive on average six citations less

than papers with gender information) once the publication year and number of authors are controlled for. This means that missing gender information is only slightly related to citation counts independently of publication year and number of authors.

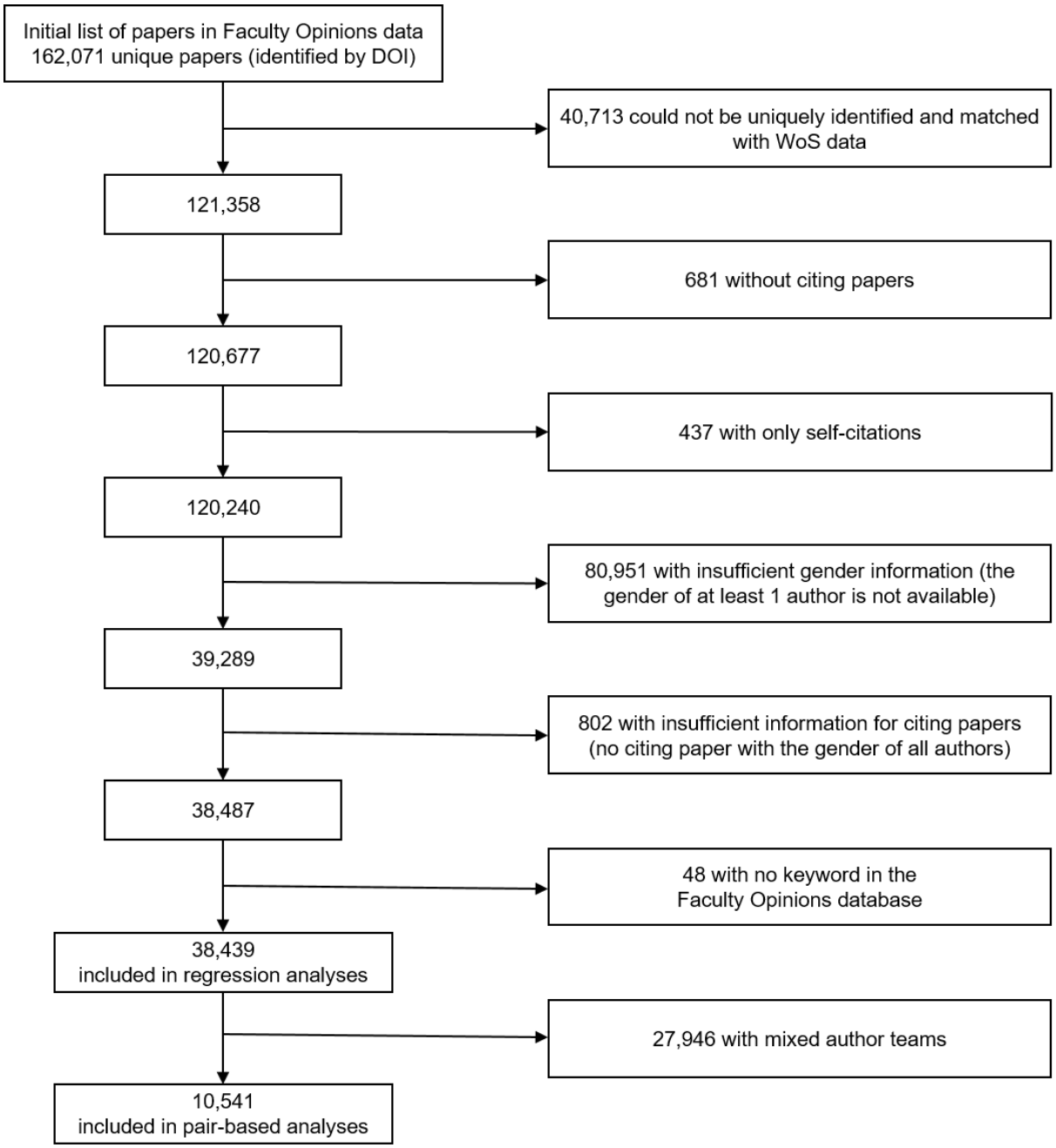


Figure 2-6. Number of papers included in the main analyses.

In our regression analyses (models M1-M3), we compared male- and mixed-authored with female-authored focal papers. Including also mixed-authored papers is one advantage of this approach (besides the possibility to use control variables). For mixed-authored papers, the share of male-authored citing papers is larger than for female-authored papers, but smaller than for male-authored papers; thus, dropping mixed-authored papers (as has been done in other analyses) yields an upper bound of gender effects on citations. We were able to include 38,439 papers in our analyses (few cases were lost due to missing information on the Faculty Opinions

keywords that we used as controls, see Figure 2-6). We controlled for the Faculty Opinions keywords by single dummy variables. These dummies measure topic similarity only on a low level of granularity: we controlled for single keywords, but not for idiosyncratic combinations of keywords that would define research topics more precisely (but see 2.5.2 for some robustness analyses with regressions based on pairs of papers that allow for more fine-grained controls of papers' similarity).

Therefore, we complemented regressions including single dummy variables with a novel approach based on pairs of papers. The alternative approach allowed us to control the papers' topic not only at a more fine grained level, but also at different levels of granularity. For these analyses presented in Figure 2-4 and Figure 2-5, we paired papers of female- and male-authored papers. We could not include mixed-authored focal papers in the analyses. While pairing the focal papers increases the number of observations (different pairs of papers are observed instead of single papers), it reduces the number of focal papers that we could consider due to the necessary omission of mixed-authored papers. From the 10,541 papers that we included in the analyses based on pairs of papers, 1,261 have only female authors and 9,280 only male authors. Building all possible pairs of papers such that one paper is authored only by female scientists and the other paper is authored only by male scientists results in 11,702,080 pairs that we could use for plotting the histograms in Figure 2-4.

In the analysis presented in Figure 2-5F based on WoS data, we considered all papers that are included in the WoS, have been published in 2012, and are of document type 'article' or 'review' (to include only substantial papers). This comprises 1,396,207 papers in total. Due to missing metadata, the set reduces to 1,235,021 papers. For 1,212,097 of the papers, the title and abstract are available in the WoS. The gender of all authors could be classified for 399,319 papers (117,143 with only female authors and 282,176 with only male authors). When matching the pairs, we only considered pairs for which both papers are in the same WoS subject category. This drastically reduces the computational complexity and makes the results more comparable to those based on the Faculty Opinions data: the papers in the Faculty Opinions database are field-specifically restricted (to biomedicine). We were able to build 508,941,740 pairs of papers with one being female-authored and the other being male-authored.

Determining gender of names

In order to infer the authors' gender (of both focal papers and their citing papers), we used an open source application that makes it possible to assign the gender to first names (Studer, 2012). This application is based on a list of 44,568 first names that have been mapped to a gender, depending on the country of origin. In order to consider the country of origin when inferring the authors' gender, we used the authors' affiliation as a proxy. In the case of unisex names, the application returns no gender to a given first name (in combination with a country of origin): the name may refer to female, male, or non-binary persons. If a first name is usually associated with a particular gender in a country, but is also used for the other gender in another country,

the application classifies this name as probably female/male. These cases are included in our analyses, i.e. a probably female (male) author is assumed to actually be female (male). Figure 2-11 shows the results for the pair-based analyses using the Faculty Opinions data when probably female/male authors are excluded and only the more reliable gender assignments are used. Since these results do not differ substantially from the results obtained when including the less reliable gender assignments, we conclude that both approaches can be used interchangeably for our analyses.

In order to operationalize the gender of author teams, we distinguished three cases: all authors are female, all authors are male, and the authors are of mixed gender. If we could not infer the gender of at least one author (because it is a unisex name or the name is not in the application's database), the paper is not included in our analyses. If multiple affiliations in different countries are linked to an author, we determined the gender of the author separately for each affiliation. If the gender classifications match, the paper remains in the analyses; if there are inconsistencies across the different classifications, the paper is not included in the analyses.

Regression analyses

We tested the regression models presented in Figure 2-2 and Table 2-1 for heteroscedasticity, multicollinearity, and outliers. To test for heteroscedasticity, we performed Breusch-Pagan tests (Breusch & Pagan, 1979). The tests are statistically significant on the 0.001 level for all three models, indicating that the variance of the error terms depends on the values of the independent variables. Therefore, we used robust standard errors for calculating p values in Table 2-1 and confidence intervals in Figure 2-2. To test for multicollinearity, we calculated the variance inflation factor (VIF) for the independent variables in all models. For the dummy variables indicating the gender of the focal papers' authors, we calculated the generalized VIF (GVIF) (Fox & Monette, 1992). GVIF accounts for the fact that the two dummy variables represent the same characteristic. Since the VIF/GVIF is smaller than five in every case, we assume a negligible level of multicollinearity in the models (James et al., 2013). To test for outliers in our data, we calculated Cook's distance for all observations. Since all values are smaller than 0.5, we do not assume any problematic effects of outliers in our analyses (Cook & Weisberg, 1982).

Similarity based on titles and abstracts

The similarity between two papers based on their titles and abstracts was calculated based on the term frequency inverse document frequency (*tf-idf*) of the words (terms) occurring in the abstracts and titles (documents). The *tf-idf* is a standard approach to obtain vector representations for documents. They indicate the relevance of each document's word in the collection of all documents (Sammut & Webb, 2010). We used the R package *text2vec* for calculating the *tf-idf*. For each paper, title and abstract were simply concatenated and the list of English stopwords provided by the R package *stopwords* was excluded. Stop-words are the most common words in the English language, such as "the," "is," "which" etc. These words do not add substantial meaning to a text and therefore should be filtered out before text mining procedures.

Furthermore, we excluded very infrequent and very frequent words to remove noise. If a word occurs less than three times over all documents, or if the proportion of documents including the word is larger than 0.3, we excluded the word. The *tf-idf* for a word *t* occurring in a document *d* of the document set *D* is defined as follows:

$$tf\text{-}idf(t,d,D) = tf(t,d) \cdot idf(d,D)$$

with

$$tf(t,d) = \frac{number\ of\ occurences\ of\ t\ in\ d}{length\ of\ d}$$

$$idf(d,D) = \log(\frac{number\ of\ documents\ in\ D}{number\ of\ Documents\ in\ D\ containing\ t})$$

tf(t,d) measures how relevant a word *t* is in document *d* (in our case: the abstract and title of a single paper). This is weighted (multiplied) by the rarity of the word in the full set of documents *D* (in our case: the pooled abstracts and titles of all papers included in the analyses). This rarity is measured in the second term by the logarithmized inverse frequency of documents in *D* which contain the word *t*.

To measure the similarity between two papers, the cosine similarity between their *tf-idf* was used. Cut-off values were needed to define different levels of similarity for which we plotted the histograms on the differences in the share of male-authored citing papers between male- and female-authored focal papers. We defined these cut-off values in a way that maximizes the comparability with the approach based on the keywords from the Faculty Opinions database to define similarity. To achieve this, we set the cut-off values so that the share of pairs that are classified as similar at each level is equal to the share of pairs that are classified as similar at each level when using the Faculty Opinions keywords to define similarity. For instance, at the first level of cosine similarity, 6.5% of pairs are included, because 6.5% of pairs have one shared Faculty Opinions keyword.

2.5.2 Supplementary text

Operationalization and results of other studies

Existing studies on gender homophily in citations used different methods and datasets. Table 2-2 summarizes these studies and shows the main differences with regard to the data and methods used, as well as their results summarized in the gender homophily rate.

Some but not all studies controlled for self-citations by excluding them in their analyses. To operationalize the authors' gender on the paper level, different methods were applied: while some studies considered all authors, others only considered the first author, the first *X* authors, and/or the corresponding authors. Mcelhinny et al. (2003) did not operationalize the authors' gender on the paper level, but analyzed all links between authors and citing authors. The gender

homophily rate was generally calculated as the difference in the share of female-authored cited references between female-authored focal papers and male-authored focal papers (for all studies listed in Table 2-2 the homophily measurement was based on the gender distribution in cited references instead of the citing papers, as we did in our main analyses). To achieve this measurement, the studies did not calculate the share of female-authored (or male-authored) cited references separately for each focal paper to summarize these shares in a next step, but instead pooled the cited references of all female-authored focal papers (male-authored focal papers). The gender homophily rate reported in Table 2-2 was then calculated as the share of female-authored cited references of female-authored papers minus the share of female-authored cited references of male-authored papers. Note that for studies that considered all authors of a paper to operationalize the gender on the paper level, mixed author teams neither belonged to the female-authored nor male-authored papers but were instead dropped, as was the case in our analyses based on pairing of papers. Figure 2-7 shows the gender homophily rate for all studies listed in Table 2-2.

Table 2-2. Studies empirically analyzing gender homophily in citations

Study	No. of citation links included	Selection of citing papers (number of papers / publication years / subfields)	Operationalization of gender	Controlling for subfields	Controlling for self-citations	Gender homophily rate
Ferber (1986)	2,394	118 / 1982-1983 / economics (subfield manpower, labor and population)	Female-only author teams vs. male-only author teams	Only citing papers of one field included	Self-citations excluded	0.116
Ferber (1988)	11,669	676 / 1982-1983 / economics, developmental psychology, mathematics, sociology	Female-only author teams vs. male-only author teams	Pairwise matching of papers based on fields	Self-citations excluded	0.081
Lutz (1990)	10,593	446 / 1982-1986 / anthropology	First author	Only citing papers of one field included	Self-citations excluded	0.078
Davenport and Snyder (1995)	4,951	100 / 1985-1994 / sociology	First author	Only citing papers of one field included	Not controlled	0.212
Mcelhinny et al. (2003)	16,766	Not available / 1965-2000 / sociolinguistic and linguistic anthropology	Author-author links considered	Only citing papers of one field included	Not controlled	0.133
Håkanson (2005)	23,483	1,739 / 1980-2000 / library and information science	Female-only author teams vs. male-only author teams	Only citing papers of one field included	Self-citations excluded	0.124

Ferber and Brün (2011)	3,256	238 / 2008 / economics (labor economics and general economics)	Female-only author teams vs. male-only author teams	Separate analysis for fields	Self-citations excluded	0.068
Knobloch-Westerwick and Glynn (2013)	2,958	1,020 / 1991-2005 / communication science	First author	Only citing papers of one field included	Not controlled	0.16
Mitchell et al. (2013)	3,013	57 / 2005 / political science (international relations)	Female-only author teams vs. male-only author teams	Only citing papers of one field included	Not controlled	0.214
Potthoff and Zimmermann (2017)	25,853	917 / 1970-2009 / communication science	Female-only author teams vs. male-only author teams (based on first two authors)	Only citing papers of one field included	Self-citations excluded	0.085
Dion et al. (2018)	30,066	1,938 / 2007-2016 / political science and social science methodology	Female-only author teams vs. male-only author teams (based on first five authors)	Separate analysis for fields	Not controlled	0.102
Ghiasi et al. (2018)	20,395,382	1,557,967 / 2008-2016 / no restriction of fields	First & corresponding authors (same gender)	Pairwise matching of papers based on topics (abstract and title)	Self-citations excluded	0.099

All studies restricted their data to a limited number of fields by including only particular journals and publication years, with the only exception being Ghiasi et al. (2018) who did not restrict their analyses to any particular field. Some studies used journals or journal sets also for further controlling the papers' research topics. Ghiasi et al. (2018) additionally used the focal papers' titles and abstracts by matching each female-authored paper to the most similar male-authored paper in the same issue of the same journal (using the gender of the first and corresponding authors to operationalize gender on the paper level). Although this approach considers more information to control for the papers' research and subject area than the approaches used in other studies, the ability to identify papers that are similar in their research questions/topics may be limited due to the relatively small number of papers – and thus topics – covered in a journal issue. Their approach also does not make it possible to control the similarity between papers at different levels of granularity, as we did in our analyses.

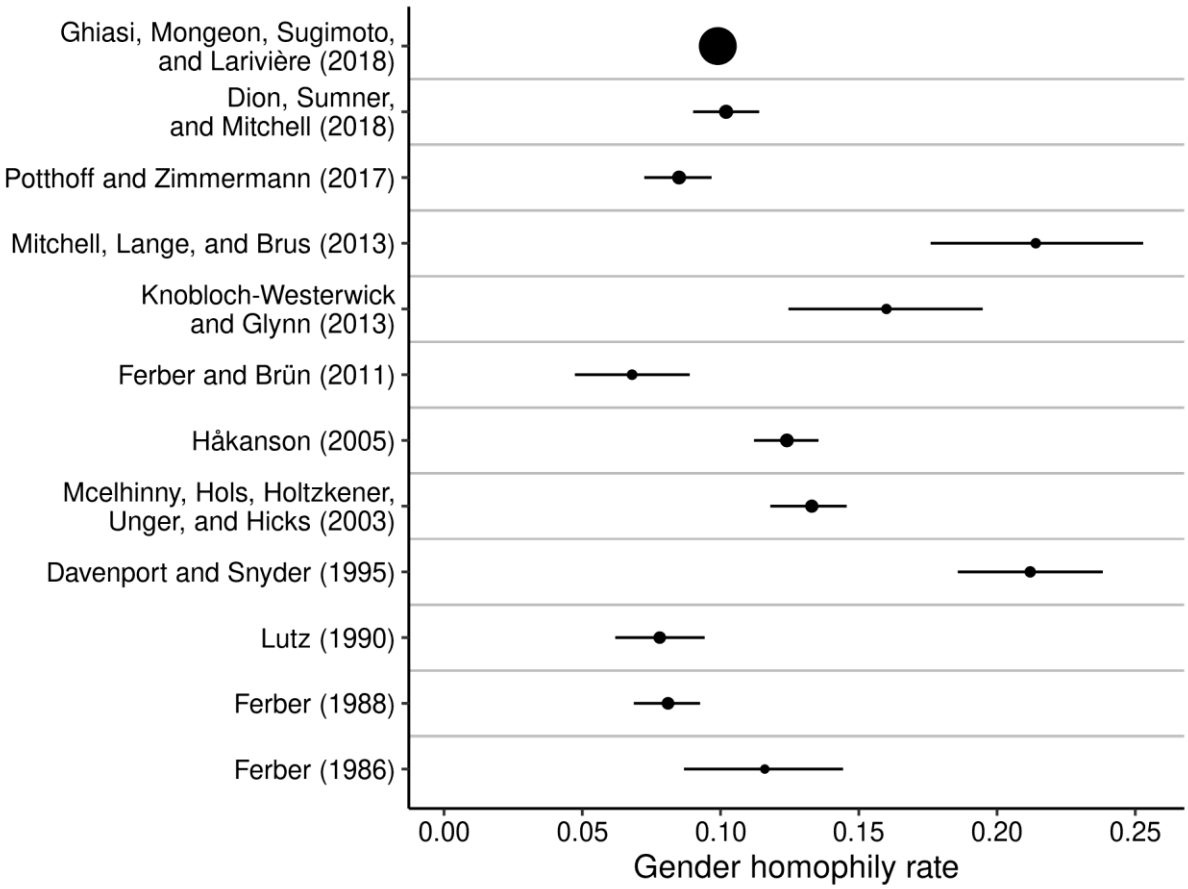


Figure 2-7. Gender homophily rates for previous studies.
Symbol sizes are proportional to the square root of the sample size. The lines show 95% confidence intervals. Studies are ordered chronologically. See Table 2-2 for details on the calculation of the gender homophily rates.

Robustness checks

In order to test the reliability of our results, we performed several robustness checks. The goal of these additional analyses was to test how the results change when different methodological approaches are applied. In general, the robustness checks support our conclusion: only a small degree of gender homophily in citations can be found after controlling for the similarity between papers, while inadequate measures of topic similarity can inflate the observed gender homophily. If not explicitly stated otherwise, all analyses described in this section are based on Faculty Opinions data.

Regression analyses using pairs of focal papers

Pairing the papers for the analyses shown in Figure 2-4 allowed us to control the papers' topic at different levels of granularity. When plotting histograms as shown in Figure 2-4, only research topics can be controlled by matching the papers at different levels of similarity. However, pairs of focal papers (instead of focal papers themselves, as in the regression analyses shown in the main text and described in 2.5.1) can also be used to perform regression analyses. In these analyses, all pairs are included where one paper is authored only by male scientists and the other paper is authored only by female scientists. The difference in the share of male-

authored citing papers is the dependent variable, and dummy variables for the number of shared Faculty Opinions keywords are the main predictors of interest. Multiple regression analyses make it possible to control not only for research topics, but also for other variables. We controlled in the regression analyses for differences in the quality ratings provided in the Faculty Opinions data (ratings of research quality provided by experts), age (publication year), and team size (number of authors) between both papers of a pair. The three variables are also included in the regression analyses on the level of focal papers presented in Figure 2-2. Table 2-3 and Figure 2-8 show the results of the regression analysis based on pairs of focal papers.

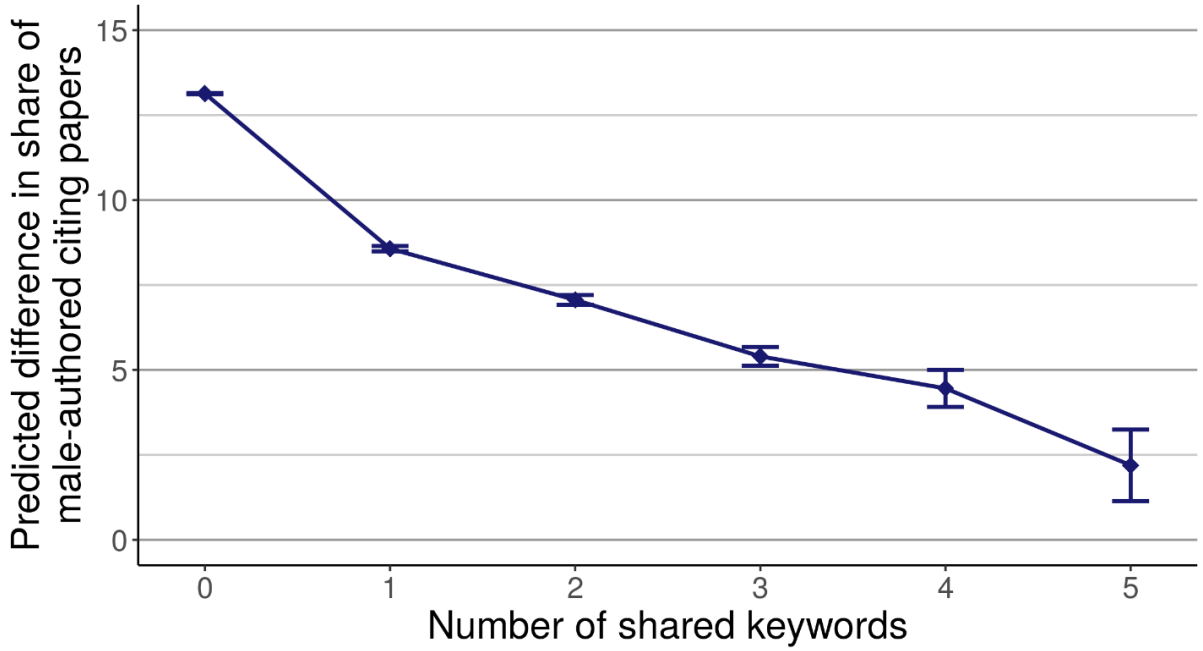


Figure 2-8. Marginal effects of the number of shared Faculty Opinions keywords.
The marginal effects are calculated based on the regression results presented in Table 2-3 for pairs of focal papers using the Faculty Opinions data. The dependent variable is the difference in the share of male-authored citing papers. Other predictor variables (difference in quality ratings, age, and team size) are set to zero.

The model includes dummy variables representing the number of shared Faculty Opinions keywords (one to four or at least five shared Faculty Opinions keywords) and the difference in the Faculty Opinions quality rating, age, and team size for a pair as independent variables. The estimated coefficients for the variables are listed in Table 2-3. Figure 2-8 reveals the predicted difference in the share of male-authored citing papers for the different numbers of shared Faculty Opinions keywords while setting the other variables (difference in quality ratings, age, and team size) to zero. Setting to zero means that the difference in the share of male-authored citing papers is predicted for the case that both papers of a pair have the same quality rating, age, and team size. The results indicate a pattern similar to the histograms in Figure 2-4: the more Faculty Opinions keywords the paired papers share (i.e., the higher their similarity in research topics), the smaller the difference in the share of male-authored citing papers gets. This confirms the result that the gendered citation patterns found in the Faculty Opinions data (in the sense that overall, male scientists are more likely to cite male-authored papers than female scientists, and

vice versa) can in large part be explained by the specialization of male and female scientists in different research topics, but not by differences in quality, age or team size. However, a high granularity of topic similarity measurement is needed (based on combinations of keywords) in order to identify the large impact of this structural aspect explaining gendered citation patterns.

Table 2-3. Regression results for pairs of focal papers (Faculty Opinions data)

	Dependent variable: difference in the share of male-authored citing papers
Number of shared Faculty Opinions keywords (reference category: 0)	
1	-4.566***
	(0.041)
2	-6.074***
	(0.073)
3	-7.737***
	(0.140)
4	-8.678***
	(0.279)
≥5	-10.942***
	(0.537)
Difference in quality ratings	2.098***
	(0.010)
Difference in age	0.375***
	(0.002)
Difference in team size	-0.665***
	(0.004)
Intercept	13.135***
	(0.009)
N	11,139,628
R^2	0.012

Note. Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Including self-citations

For the analyses described in the main text, we excluded all self-citations because self-citations artificially increase the observed degree of gender homophily. We defined self-citations as focal paper-citing paper pairs for which at least one author of the focal paper is identical to at least one author of the citing paper (Glänzel et al., 2004). This is based on the assumption that papers were written by the identical author if the focal paper and the citing paper have at least one common author name, taking into account the first initial and the full surname. It can be expected that in some cases the first initial and the full surname are shared by different persons and therefore the assumption that the authors are the same person is wrong. At the same time, it can be assumed that the name representation rarely differs for a person (Backes, 2018). Thus, this approach for identifying self-citations is likely to overestimate the existence of self-citations, but most self-citations can be assumed to be found.

Figure 2-9 shows the histograms for the differences in the share of male-authored citing papers for pairs of focal papers (similar to Figure 2-4) when including self-citations. Confirming other studies (e.g., Ghiasi et al., 2018; Håkanson, 2005; Lutz, 1990), these results show that self-citations contribute to gendered citation patterns. The pattern that the difference in the share of male-authored citing papers becomes smaller the better the topic is controlled (i.e., the minimum number of shared keywords increases) does not change when self-citations are included in the analyses. At the same time, these extended analyses suggest that evidence for gender homophily is overestimated when self-citations are not excluded (as was done in some previous studies, see Table 2-2). In order to validly measure gender homophily as a preference for citing colleagues of the same gender (and not just one's own work), our empirical evidence clearly reveals that this methodological step of excluding self-citations is very important. There is a much larger average difference in the share of citing papers authored by males left that one might erroneously interpret as gender homophily when not excluding self-citations: the difference increases by 7.2 percentage points when matching papers with at least five shared Faculty Opinions keywords (the difference is then 8.83, instead of 1.61 when excluding self-citations; see Figure 2-4). The lager difference that is driven by self-citations should not be confused with homophily.

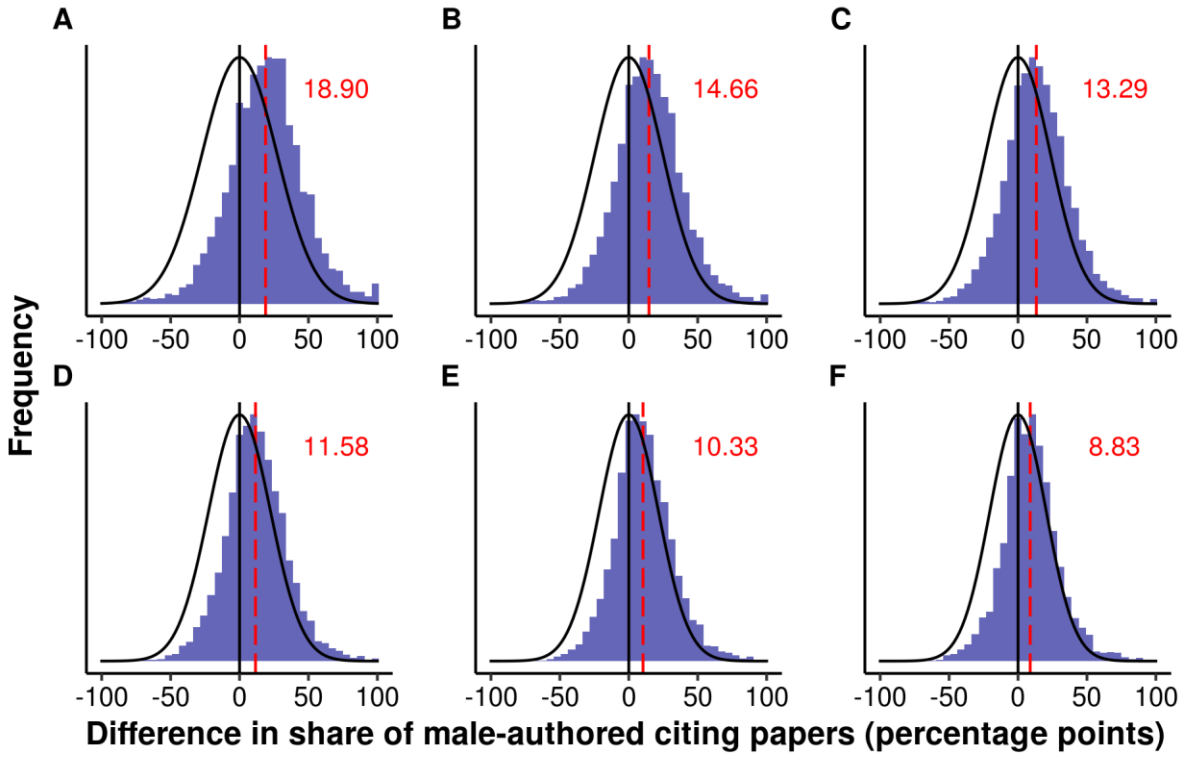


Figure 2-9. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, including self-citations).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citations than the female-authored paper of this pair. The histograms differ in the minimum number of shared Faculty Opinions keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,932,148 pairs in (A), 779,244 pairs with at least one shared Faculty Opinions keyword in (B), 227,093 pairs with at least two shared Faculty Opinions keywords in (C), 59,122 pairs with at least three shared Faculty Opinions keywords in (D), 14,305 pairs with at least four shared Faculty Opinions keywords in (E), and 3,035 pairs with at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Using cited references instead of citing papers

Most of the existing studies on gender homophily in citations analyzed references cited in focal papers in order to answer the question whether female-authored focal papers are less likely to refer to male-authored papers compared to male-authored focal papers. Rather than following this approach of analyzing the cited references in focal papers, we decided to use the papers that cited the focal papers in order to analyze the question whether male-authored focal papers are more likely to be cited by male scientists than female-authored focal papers. The major advantage of using citing papers (instead of cited references) is that the publication year of the papers (which are used to measure homophily bias) can be held constant to a greater extent in the analyses. The approach allows for a better standardization of the overall gender composition in science, which might influence gender-specific citation patterns. The gender distribution changed over time, and this time-trend may lead to an overestimation of homophily bias: the

analysis of cited references can go a long way back in time when the share of female scientists who could be cited was smaller. Older papers written predominantly by men are more likely to be cited by men, simply because they are on average more senior. Senior researchers may work on more classical topics with references reaching longer back in time than junior researchers. Since the focal papers we studied were published no earlier than 2002, the papers citing them could also not have been published before 2002 either. This means that the risk pool of citing papers covers a much smaller time frame than in the analysis of cited references. However, to produce results that are better comparable with most other studies on gender homophily in citations, we considered cited references instead of citing papers in a robustness check. Figure 2-10 shows the histograms for the differences in the share of male-authored cited references for pairs of focal papers.

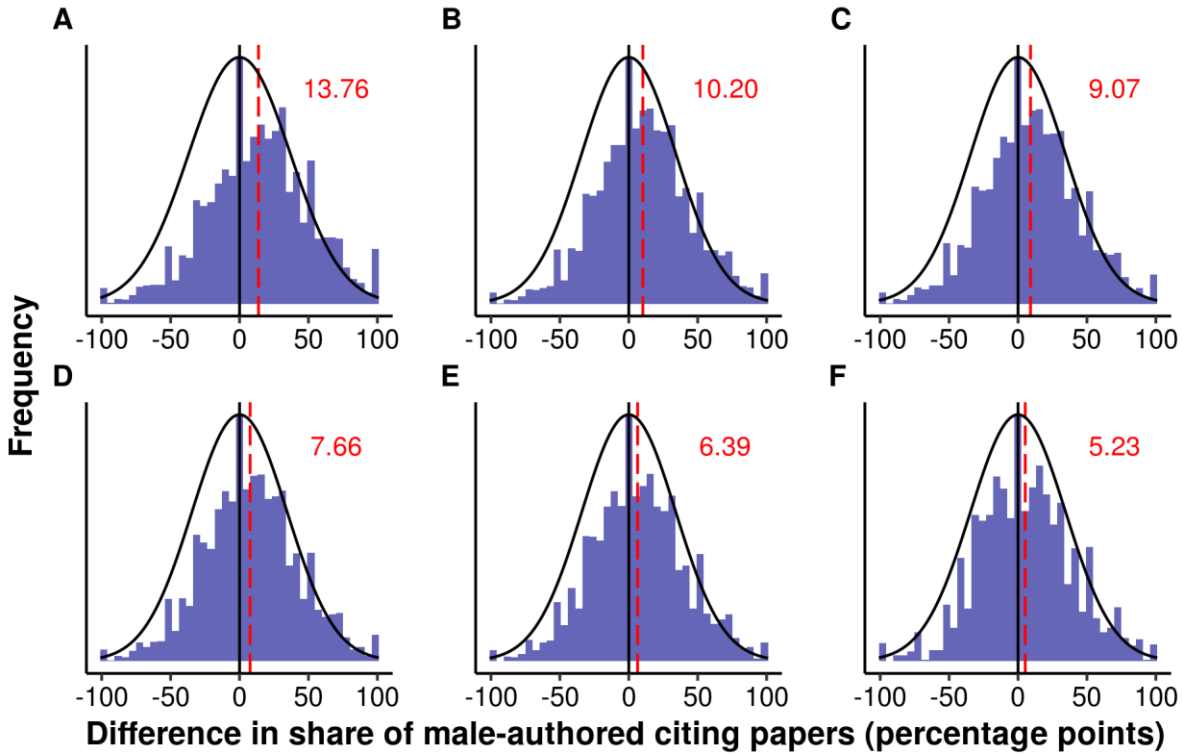


Figure 2-10. Histograms for the differences in the share of male-authored cited references for pairs of focal papers (Faculty Opinions data).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored cited references than the female-authored paper of this pair. The histograms differ in the minimum number of shared Faculty Opinions keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 9,654,680 pairs in (A), 633,651 pairs with at least one shared Faculty Opinions keyword in (B), 182,092 pairs with at least two shared Faculty Opinions keywords in (C), 46,830 pairs with at least three shared Faculty Opinions keywords in (D), 11,005 pairs with at least four shared Faculty Opinions keywords in (E), and 2,241 pairs with at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Similar to the results shown in Figure 2-4, the average difference in the share of male-authored cited references (our indicator for homophily) decreases when topic similarity is controlled for more thoroughly. But when analyzing cited references instead of citing papers, the difference does not diminish as much: in the analysis of cited references, the difference in the share of papers authored by males decreases from 13.76 to 5.23 percentage points, while the difference in the share of male-authored citing papers (the approach we took for our main analysis) diminishes from 12.64 to 1.61 percentage points (see Figure 2-4). The difference between both results suggests that controlling publication year in the analysis is important in order to validly measure gender homophily in citations. In the analysis of citing papers, the publication year of the papers can be hold constant to a greater extent than in the analysis of cited references (see above).

Gender assignments

For a given first name and country (if available), the database that we used for inferring the authors' gender differentiates between "is mostly female/male" and "is female/male". In our main analyses, we used both types of gender classification. Figure 2-11 presents empirical results as shown in Figure 2-4 but dropping all names with less reliable gender classifications: in the histograms, the difference in the share of male-authored citing papers for pairs of focal papers is shown only for papers with the more reliable gender assignment "is female/male." The results are very similar to using both types of gender classification: the difference in the share of male-authored citing papers decreases from 13.58 to 0.62 percentage points when using the more restrictive gender assignment, while it diminishes from 12.64 to 1.61 percentage points when using the less restrictive assignments including also the "mostly female/male" classification (see again Figure 2-4 in our main analyses). We conclude that the reliability of gender assignments indicated by the database does not play a significant role for our results.

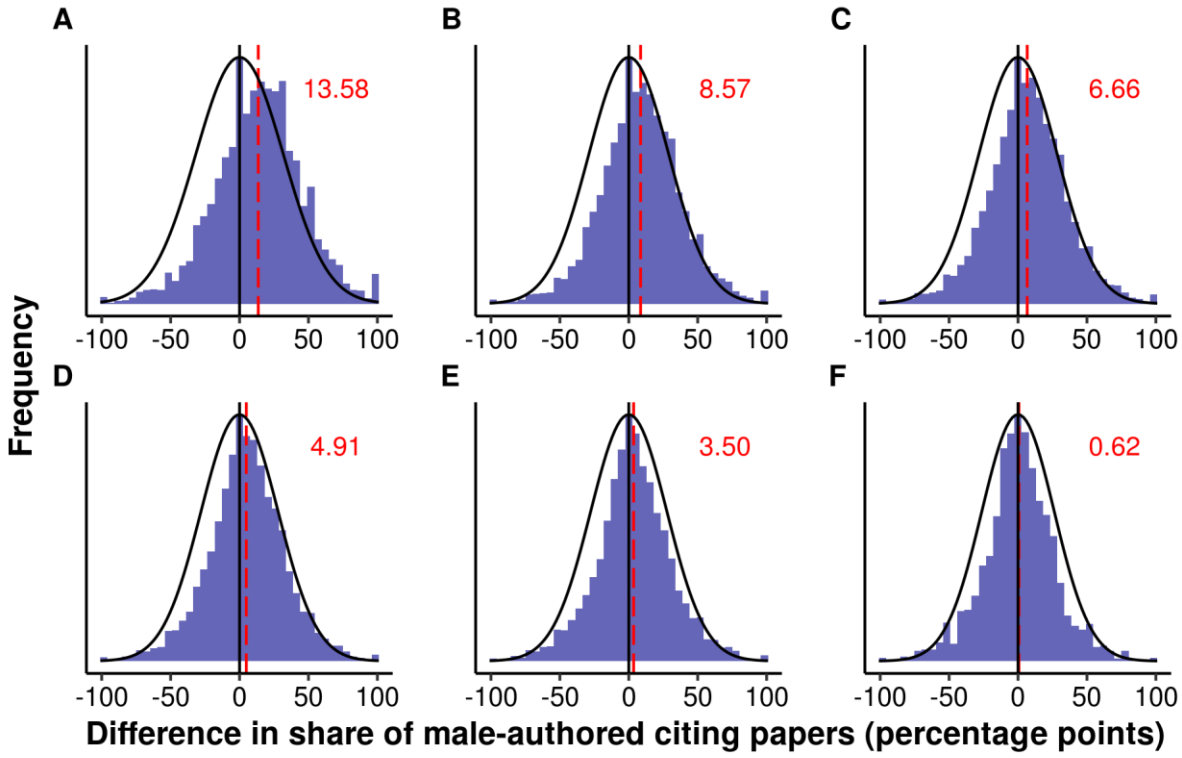


Figure 2-11. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, based on more restrictive gender assignments).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum number of shared Faculty Opinions keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 7,532,670 pairs in (A), 514,725 pairs with at least one shared Faculty Opinions keyword in (B), 150,964 pairs with at least two shared Faculty Opinions keywords in (C), 39,729 pairs with at least three shared Faculty Opinions keywords in (D), 9,818 pairs with at least four shared Faculty Opinions keywords in (E), and 2,107 pairs with at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Alternative approaches for controlling similarity

Similarity between papers can be controlled by various approaches. We deem abstracts and titles the most adequate alternative to Faculty Opinions keywords (when expanding the results to papers not included in the Faculty Opinions database). Abstracts and titles usually contain comprehensive information about the content and research topics of papers. Figure 2-5B and Figure 2-5F in the main text show average differences in the share of male-authored citing papers when using abstracts and titles to match the papers. Figure 2-12 and Figure 2-13 present the corresponding histograms based on this alternative approach for both Faculty Opinions and WoS data. Besides using titles and abstracts, other possibilities for controlling similarity between papers are the number of shared cited references, the number of shared keywords provided in the WoS, or the number of shared WoS subject categories (Thijs, 2019). WoS

keywords include both keywords specified by a paper's authors and keywords automatically generated based on the titles of a paper's cited references (Garfield, 1990; Garfield & Sher, 1993). Figure 2-14, Figure 2-15, and Figure 2-16 show the results for these other approaches using the Faculty Opinions data. When interpreting the results in the figures, it should be considered that the similarity levels do not align with the similarity levels based on the Faculty Opinions keywords; i.e., there is a lower or higher number of paper pairs that are categorized into a certain similarity level than when using the keyword approach. This is because the number of shared cited references, WoS keywords, and WoS subject categories are discrete values: such values cannot be perfectly recategorised into different similarity levels (shares of papers to be found to be similar) that result when using different numbers of Faculty Opinions keywords to define similarity levels. Replication of the percentage distribution across different similarity categories based on Faculty Opinions keywords is only possible with the measurement of similarity based on abstracts and titles. The cosine similarity is a continuous measure that can be categorized at arbitrary cut-off values.

Although we used very different approaches for measuring paper similarity, we always found the same pattern: the better the topic similarity between papers is controlled for, the smaller the difference in the share of male-authored citing papers gets. Since the approaches differ with regard to the share of paper pairs that fall into different similarity levels, the approaches cannot be directly compared. However, one might argue that different similarity levels of two approaches are roughly comparable if their numbers of pairs are approximately equal. For example, the number of pairs with at least one shared cited reference (23,192) roughly corresponds to the number of pairs that share at least four Faculty Opinions keywords (14,167). Thus, comparing these similarity levels of the two approaches means comparing the 23,192 most similar paper pairs when measuring the similarity based on shared cited references with the 14,167 most similar pairs when measuring the similarity based on shared Faculty Opinions keywords. For this comparison, the difference in the share of male-authored citing papers (our indicator of homophily) replicates well: it is only slightly higher when using the overlap in cited references instead of keywords to define similarity (5.88 vs. 3.48 percentage points; see Figure 2-14). Likewise, three shared cited references (resulting in 2,354 pairs that could be matched) roughly correspond to five shared Faculty Opinions keywords (resulting in 3,010 pairs that could be matched). At this level of similarity, the difference in the share of male-authored citing papers is also only slightly higher when using cited references for measuring topic similarity (2.75 vs. 1.61 percentage points).

The effect of controlling WoS keywords is comparable to the effect of controlling similarity based on titles and abstracts for the Faculty Opinions data. The difference in the share of male-authored citing papers decreases from 12.64 percentage points when including all pairs of papers to 7.27 when including the 2,312 pairs of papers with at least three shared WoS keywords (see Figure 2-15). For the approach based on abstracts and titles, the difference in the share of male-authored citing papers decreases from 12.76 to 6.41 percentage points at the highest level

of similarity, where 2,707 papers could be matched. This decrease is also very similar to the one observed when using at least three shared WoS subject categories (which is from 12.64 to 5.41 percentage points; see Figure 2-16). Further restricting the pairs of focal papers to those with at least four or five shared WoS subject categories increases the difference in the share of male-authored citing papers. However, these analyses are based only on a very small number of cases (126 and 44 pairs of focal papers). A small number of cases implies a limited reliability of the results.

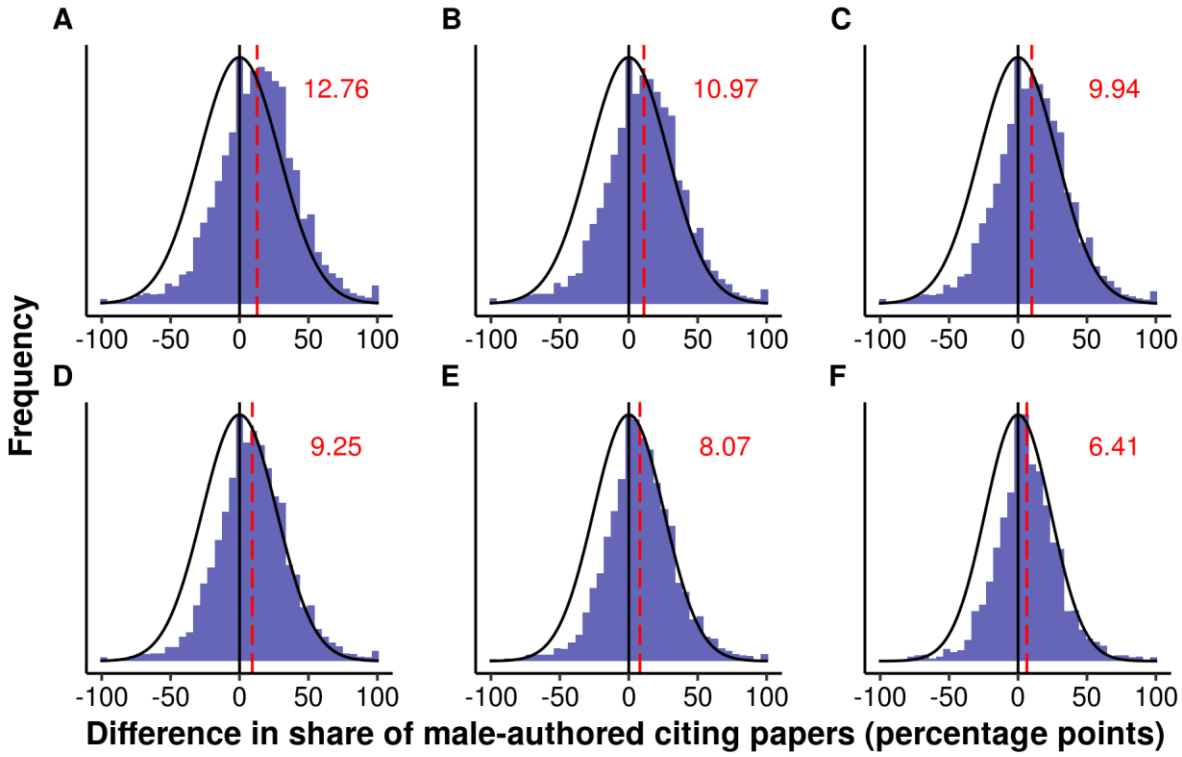


Figure 2-12. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, using titles and abstracts for measuring the similarity between papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum cosine similarity between the *tf-idf* of two paired papers, and – as a consequence – in the number of pairs of focal papers included: all 10,730,525 pairs in (A), 703,162 pairs with a cosine similarity of at least 0.026 in (B), 204,596 pairs with a cosine similarity of at least 0.047 in (C), 53,013 pairs with a cosine similarity of at least 0.084 in (D), 12,746 pairs with a cosine similarity of at least 0.147 in (E), and 2,707 pairs with a cosine similarity of at least 0.246 in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

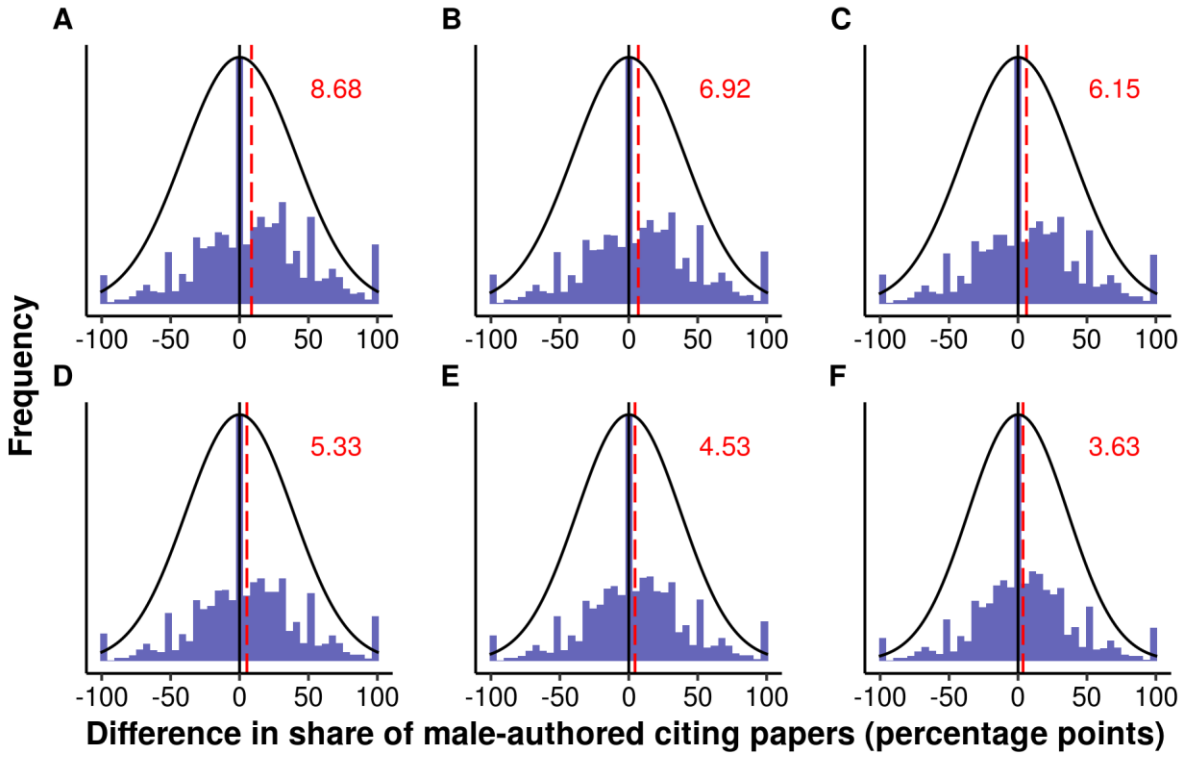


Figure 2-13. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (WoS data, using titles and abstracts for measuring the similarity between papers).

In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum cosine similarity between the *tf-idf* of two paired papers, and – as a consequence – in the number of pairs of focal papers included: all 508,941,740 pairs in (A), 32,903,466 pairs with a cosine similarity of at least 0.036 in (B), 9,572,570 pairs with a cosine similarity of at least 0.064 in (C), 2,510,050 pairs with a cosine similarity of at least 0.112 in (D), 615,604 pairs with a cosine similarity of at least 0.185 in (E), and 133,381 pairs with a cosine similarity of at least 0.286 in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

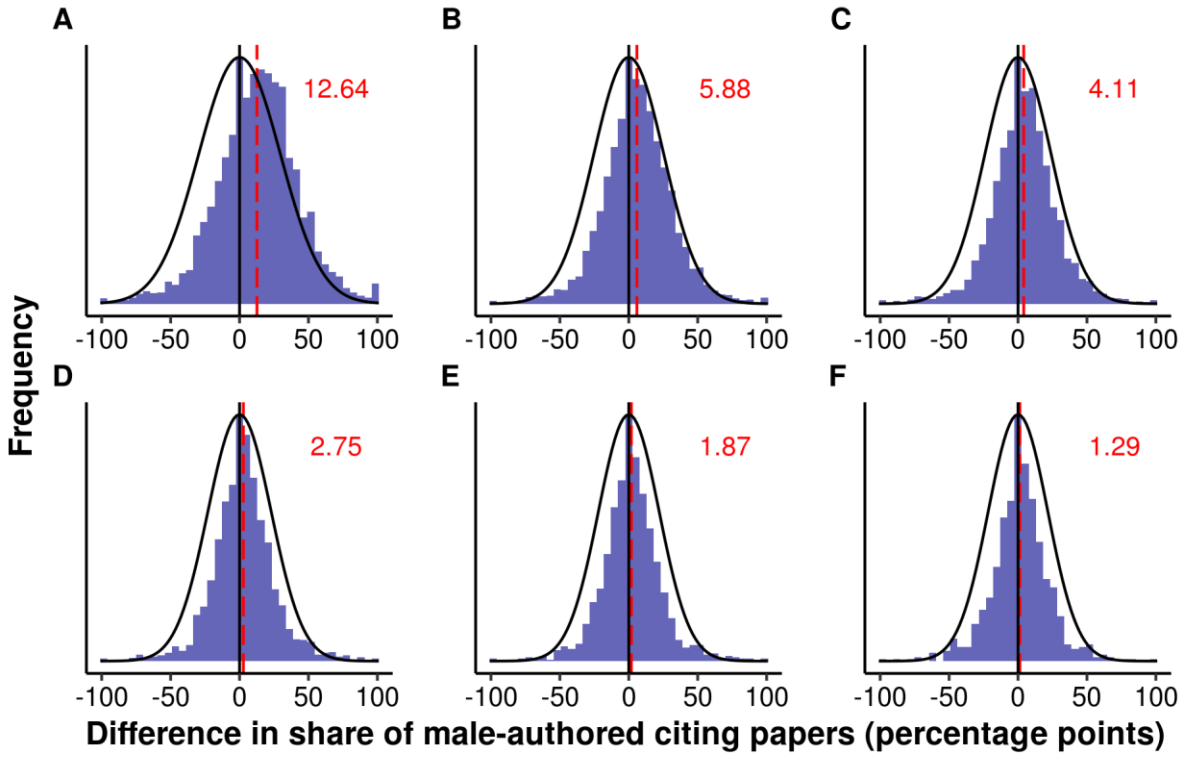


Figure 2-14. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, using the number of shared cited references for measuring the similarity between papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum number of shared cited references that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,702,080 pairs in (A), 23,192 pairs with at least one shared cited reference in (B), 5,479 pairs with at least two shared cited references in (C), 2,354 pairs with at least three shared cited references in (D), 1,304 pairs with at least four shared cited references in (E), and 798 pairs with at least five shared cited references in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

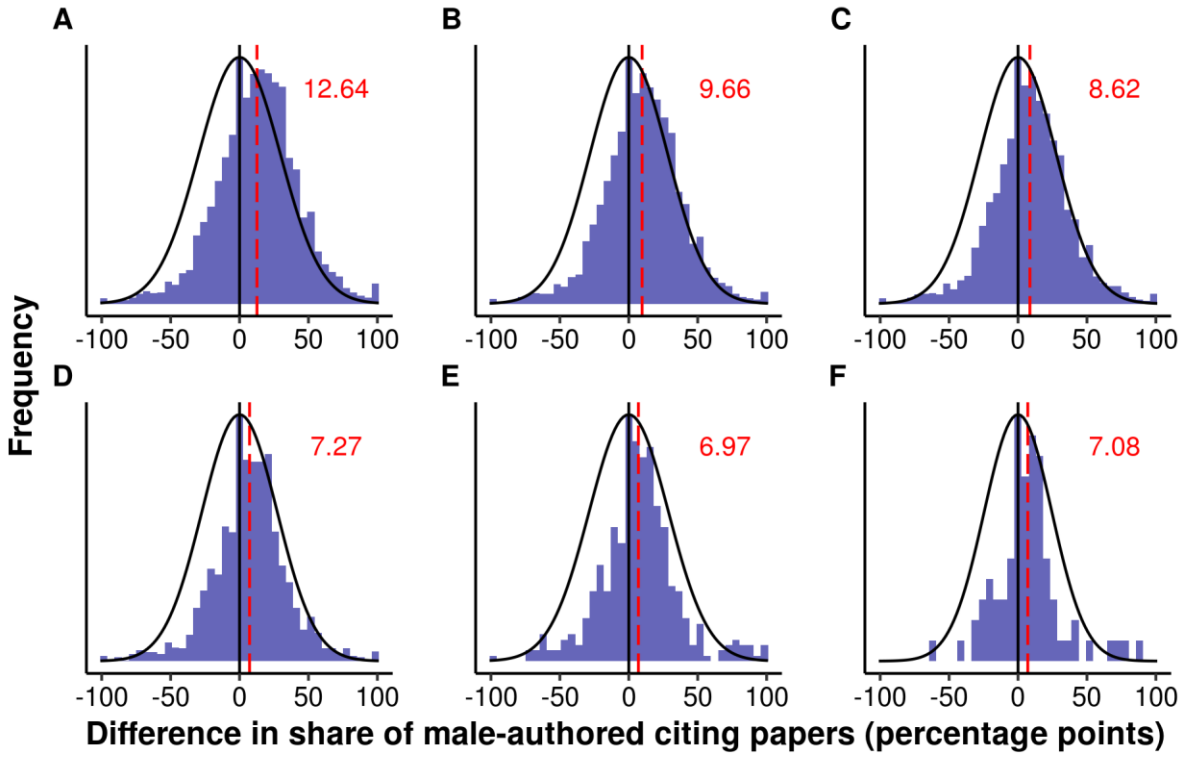


Figure 2-15. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, using the number of shared WoS keywords for measuring the similarity between papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum number of shared WoS keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,702,080 pairs in (A), 330,072 pairs with at least one shared WoS keyword in (B), 23,074 pairs with at least two shared WoS keywords in (C), 2,312 pairs with at least three shared WoS keywords in (D), 409 pairs with at least four shared WoS keywords in (E), and 81 pairs with at least five shared WoS keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

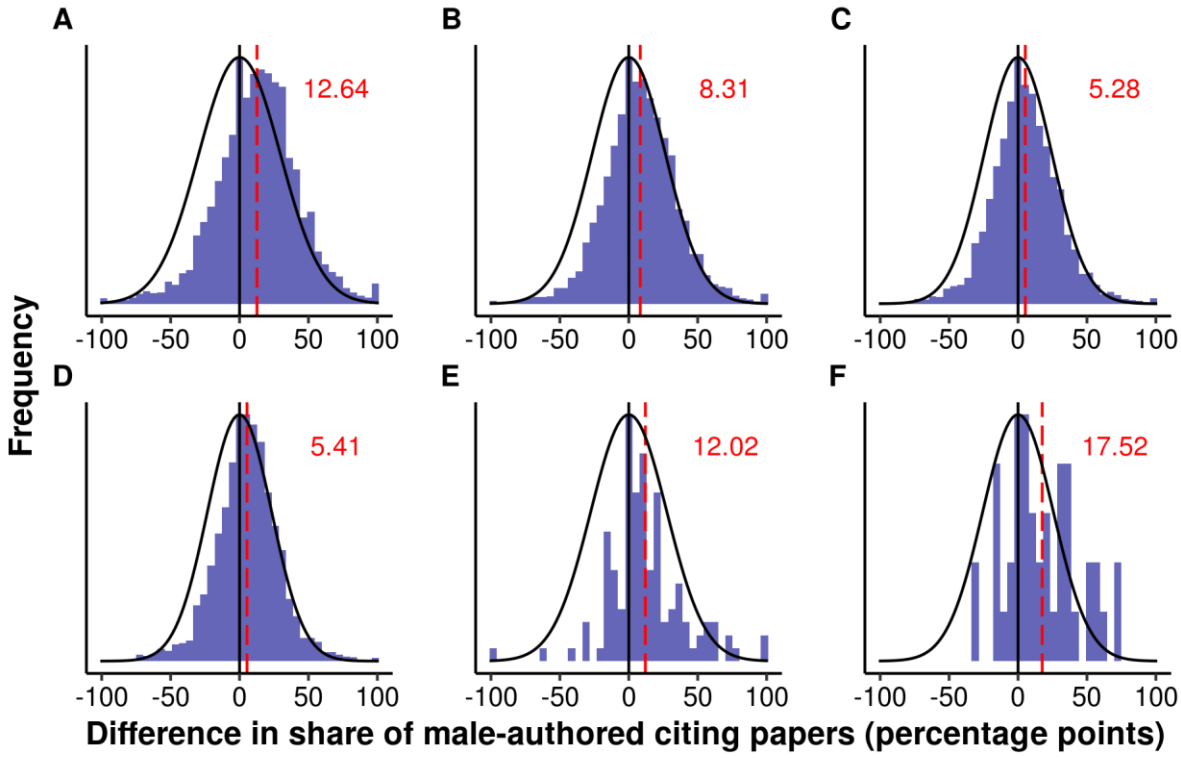


Figure 2-16. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, using the number of shared WoS subject categories for measuring the similarity between papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum number of shared subject categories that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,702,080 pairs in (A), 745,887 pairs with at least one shared subject category in (B), 56,975 pairs with at least two shared subject categories in (C), 6,180 pairs with at least three shared subject categories in (D), 126 pairs with at least four shared subject categories in (E), and 44 pairs with at least five shared subject categories in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

All in all, our extensions to other measures of similarity indicate a strong robustness of our main conclusion: topic similarity is an important structural aspect that should be controlled when one is interested in direct gender effects, such as gender homophily bias. Only when the specialization of male and female scientists on research topics is controlled for with exact measurements of the different research topics and questions they work on, one can see the "bias" that leads same gender peers to cite each other more often than what a baseline model of gender-blind selection of relevant literature would predict. A noteworthy side-result of our extensions to other measures is their evaluation in regard to their reliability in capturing topic similarity. It appears that expert ratings based on standardized lists of keywords are more suitable in defining risk pools of papers that belong to a common research field/topic than measurements based on alternative indicators of papers' content. Similarity measures based on abstracts or titles, cited references or WoS keywords may offer viable alternative measures of topic similarity when

expert ratings are not available (as in the Faculty Opinions data). However, their disadvantage is that only a small number of pairs with (several) shared cited references or WoS keywords can be matched. Titles and abstracts allow for a more nuanced measurement of topic similarity, since they allow defining different levels of similarity with a considerable number of pairs that can be matched. The key take-away of these robustness checks is, however, that thorough controls for research topics are important for identifying genuine gender effects.

Excluding papers with extreme gender distributions among citing papers

Papers included in the Faculty Opinions data are on average cited more often than other papers in the same field (Waltman & Costas, 2014). Thus, there are more focal papers with large citation counts in the Faculty Opinions data (that we used in the main analyses in this study) than in the WoS data. Having smaller citation counts in the WoS data increases the chance of having 100% female-authored or 100% male-authored citing papers, because there are fewer possible shares of female-authored and male-authored citing papers. For example, a paper with one citation can only have 0% or 100% male-authored (female-authored) citing papers, a paper with two citations can only have 0%, 50% or 100% male-authored (female-authored) citing papers (again, excluding mixed-authored papers, to be able to pair only male- and female-authored papers). While papers with large citation counts may have a share of male-authored (female-authored) citing papers close to but less than 100%, most papers with few citation counts would have a share of male-authored (female-authored) citing papers of exactly 100% in a similar situation. For the pairs of focal papers, this increases the chance of having no difference in the share of male-authored citing papers (which may be a reason for the large number of pairs with a difference of zero in Figure 2-13). However, if there is a gender homophily bias in citation decisions, the average difference in the share of male-authored citing papers on the aggregated level should still be larger than zero for the focal papers with only one (a few) citation(s). In order to test whether the large share of pairs of focal papers with no difference in the share of male-authored citing papers affects our results, we generated the histograms for the pairs of focal papers after excluding all papers with 100% female-authored or 100% male-authored citing papers. We did this for the Faculty Opinions and WoS data, using abstracts and titles for measuring the similarity between papers in both cases. In either case, including and excluding papers with only female-authored or only male-authored citing papers produces almost identical results. The share of male-authored citing papers reduces from 12.76 to 6.41 percentage points (Faculty Opinions data; see Figure 2-12) and from 8.68 to 3.63 percentage points (WoS data; see Figure 2-13) when including papers with extreme gender distributions among citing papers (100% male-authored or female-authored papers). When excluding them, the difference in the share of male-authored citing papers decreases from 12.99 to 6.52 percentage points (Faculty Opinions data; see Figure 2-17) and from 9.70 to 4.02 percentage points (WoS data; see Figure 2-18). Thus, we observe the same pattern as in other analyses: controlling for the similarity between papers reduces the difference in the share of male-authored citing papers as evidence

for gender homophily; and this result is very robust to using alternative sample restrictions (here: including and excluding papers with extreme gender distributions among citing papers).

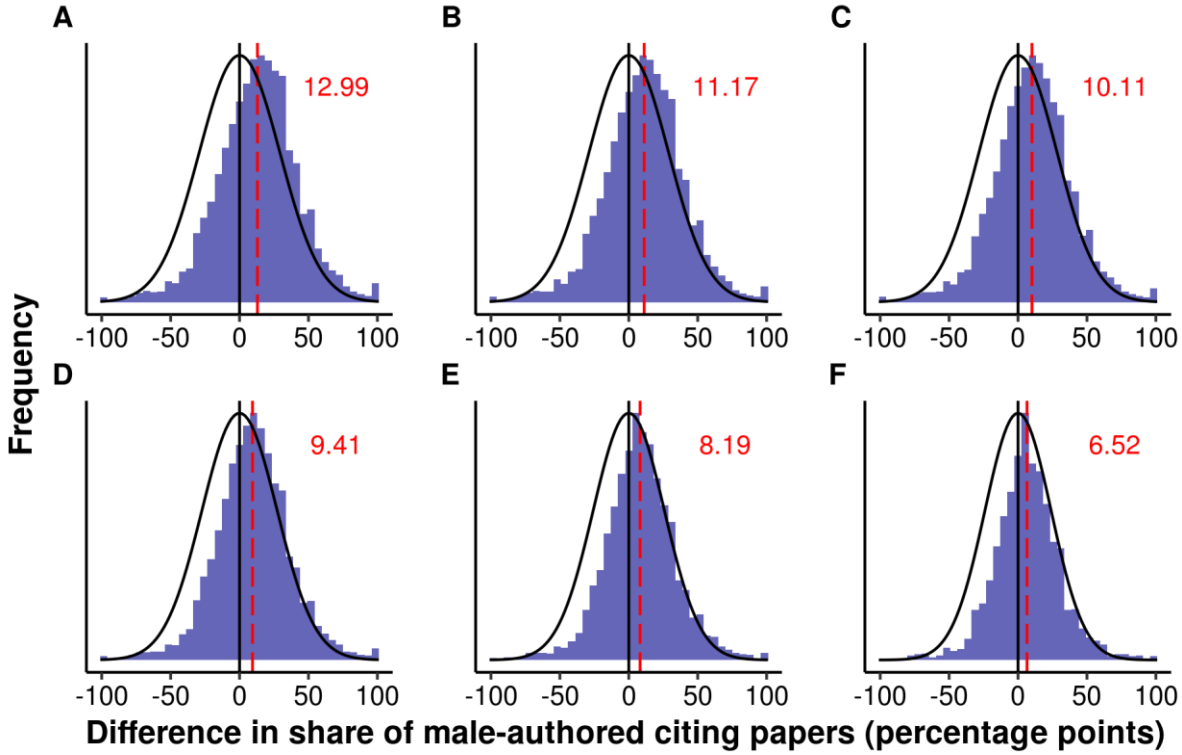


Figure 2-17. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (Faculty Opinions data, using titles and abstracts for measuring paper similarity and excluding papers with extreme gender distributions among citing papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum cosine similarity between the *tf-idf* of two paired papers, and – as a consequence – in the number of pairs of focal papers included: all 10,544,969 pairs in (A), 689,035 pairs with a cosine similarity of at least 0.026 in (B), 200,286 pairs with a cosine similarity of at least 0.047 in (C), 51,895 pairs with a cosine similarity of at least 0.084 in (D), 12,486 pairs with a cosine similarity of at least 0.147 in (E), and 2,650 pairs with a cosine similarity of at least 0.246 in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

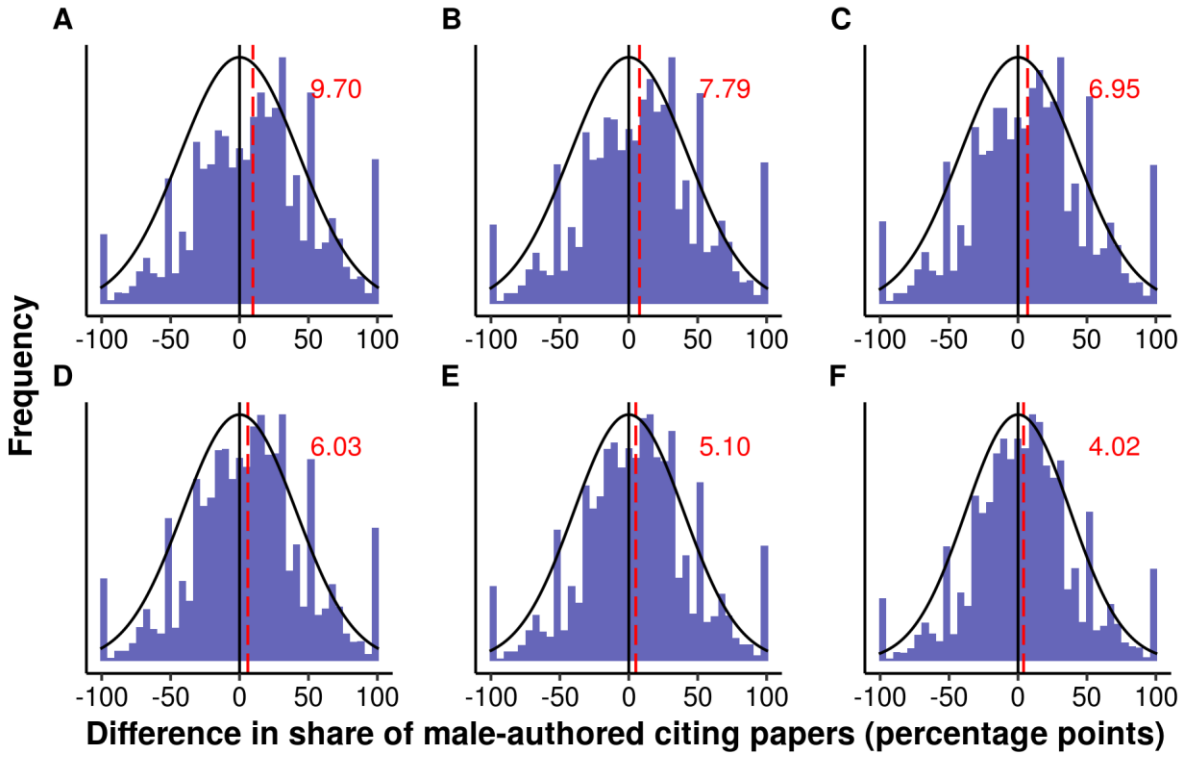


Figure 2-18. Histograms for the differences in the share of male-authored citing papers for pairs of focal papers (WoS data, using titles and abstracts for measuring paper similarity and excluding papers with extreme gender distributions among citing papers).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the male-authored paper of a pair has a higher share of male-authored citing papers than the female-authored paper of this pair. The histograms differ in the minimum cosine similarity between the *tf-idf* of two paired papers, and – as a consequence – in the number of pairs of focal papers included: all 455,434,277 pairs in (A), 29,648,614 pairs with a cosine similarity of at least 0.035 in (B), 8,631,470 pairs with a cosine similarity of at least 0.064 in (C), 2,252,018 pairs with a cosine similarity of at least 0.112 in (D), 546,588 pairs with a cosine similarity of at least 0.185 in (E), and 116,950 pairs with a cosine similarity of at least 0.287 in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Level of analysis

Using pairs of focal papers may result in certain papers having a stronger influence on the results than other papers. For example, imagine two female-authored papers: paper A, for which there are nine male-authored papers with two shared Faculty Opinions keywords, and paper B, for which there is only one male-authored paper with two shared Faculty Opinions keywords. This means that in the analysis based on all possible pairs with at least two shared Faculty Opinions keywords, nine pairs containing paper A are considered, but only one pair containing paper B. In this scenario, paper A would have a stronger influence on the result than paper B, since 90% of pairs of papers are based on paper A and only 10% on paper B. Extreme values in the share of male-authored citing papers for papers that are included in many pairs would have a great effect on the results. We assume that this should not make a difference, since

extreme values can be expected to occur at both ends of the spectrum between a small and a large share of male-authored citing papers. In order to empirically verify our assumption, we performed additional analyses by changing the level of analysis from pairs of focal papers to focal papers. For every male-authored focal paper, we calculated the average difference in the share of male-authored citing papers to all paired female-authored focal papers. This results in one value for each male-authored focal paper, which can be interpreted as the average difference in the share of male-authored citing papers between the male-authored focal paper and its paired female-authored focal papers.

Figure 2-19 shows the histograms of these values by the number of shared Faculty Opinions keywords. Figure 2-20 shows the corresponding results when aggregating the differences in the share of male-authored citing papers for each female-authored focal paper instead of each male-authored focal paper. The results of these two analyses differ only slightly from each other: the difference in the share of male-authored citing papers decreases from 12.64 to 2.70 percentage points when aggregating over male-authored focal papers and from 12.64 to 3.77 percentage points when aggregating over female-authored focal papers. Both results also differ only slightly from the results obtained when not aggregating over focal papers (recall again the result from our main analysis, in which the difference in the share of male-authored citing papers decreases from 12.64 to 1.61 percentage points). This makes us confident that the findings of our main approach of studying pairs of focal papers is not driven by some dominant "outlier" papers that have particularly strong influence on the results.

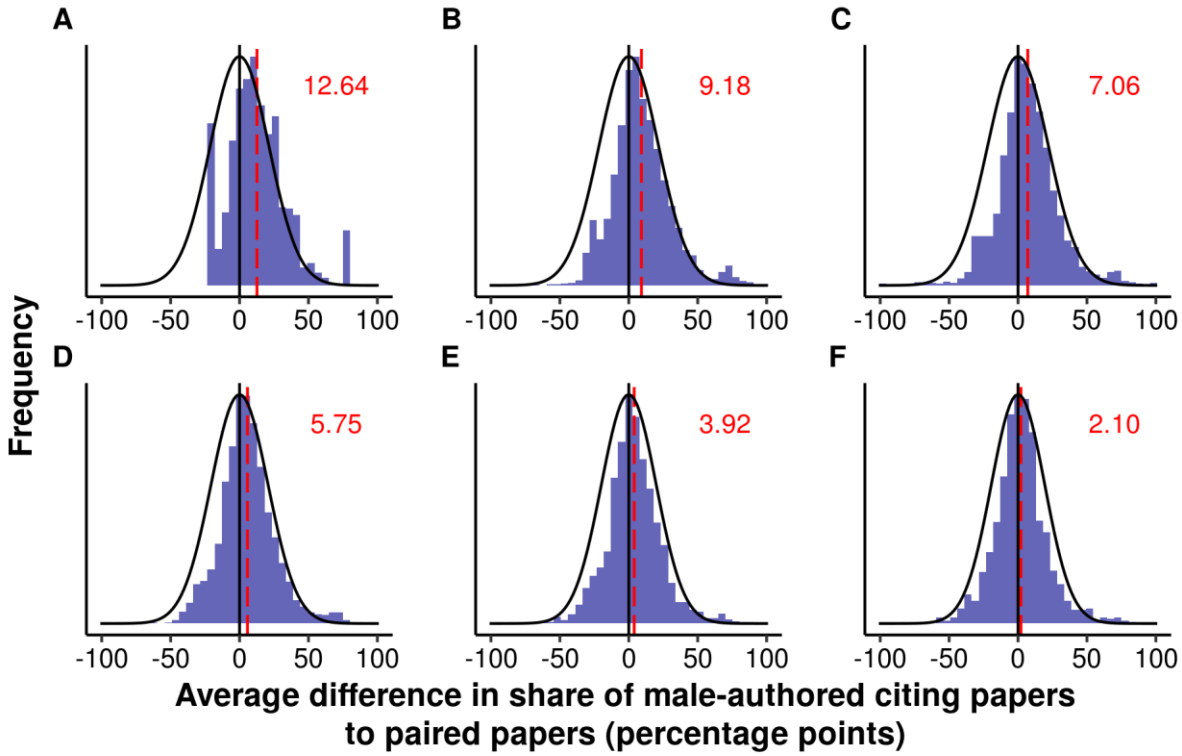


Figure 2-19. Histograms for the male-authored focal papers' average difference in the share of male-authored citing papers to their paired female-authored focal papers (Faculty Opinions data).
In each histogram, the male-authored focal papers are restricted to those that could be paired with at least one female-authored focal paper. Positive average differences result when the share of male-authored citations is higher for the male-authored focal paper than the average share for its paired female-authored papers. The histograms differ in the minimum number of shared keywords that the pairs of focal papers have, and – as a consequence – in the number of male-authored focal papers included: all 9,280 papers in (A), 9,150 papers that could be paired based on at least one shared Faculty Opinions keyword in (B), 7,533 papers that could be paired based on at least two shared Faculty Opinions keywords in (C), 5,062 papers that could be paired based on at least three shared Faculty Opinions keywords in (D), 2,644 papers that could be paired based on at least four shared Faculty Opinions keywords in (E), and 1,110 papers that could be paired based on at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

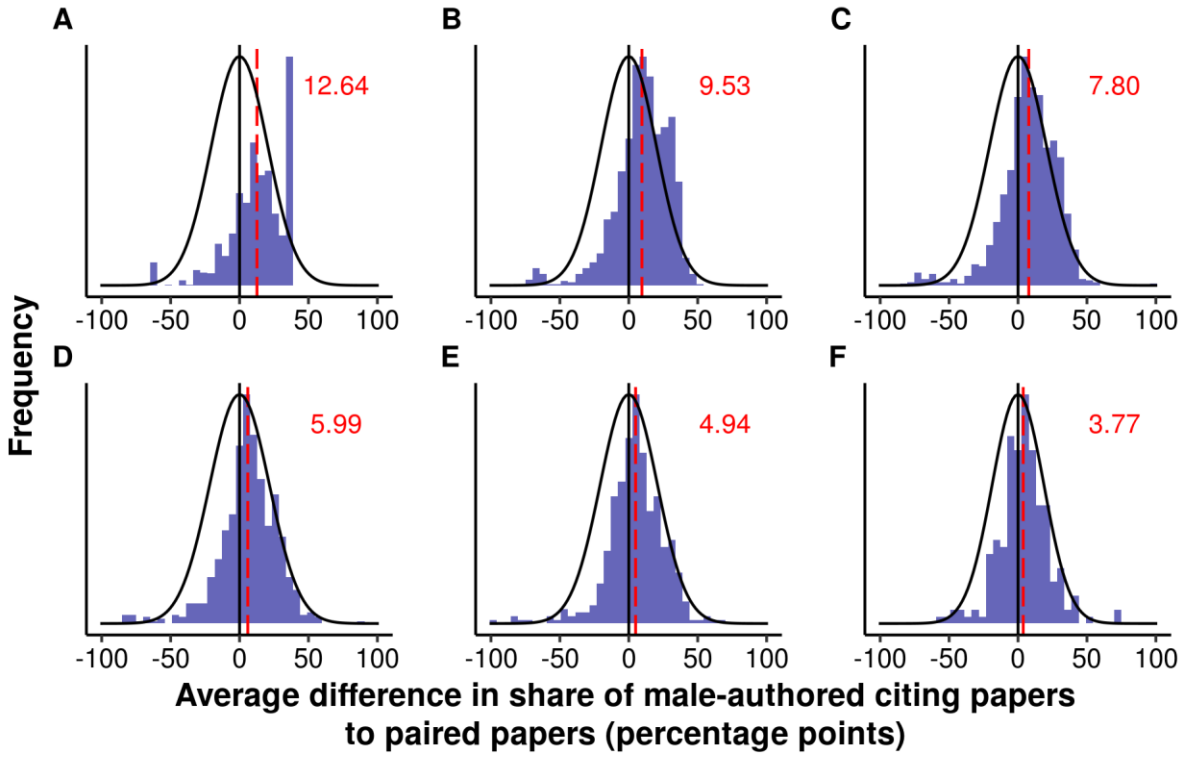


Figure 2-20. Histograms for the female-authored focal papers' average difference in the share of male-authored citing papers to their paired male-authored focal papers (Faculty Opinions data).
In each histogram, the female-authored focal papers are restricted to those that could be paired with at least one male-authored focal paper. Positive average differences result when the share of male-authored citations is smaller for the female-authored focal paper than the average share for its paired male-authored papers. The histograms differ in the minimum number of shared keywords that the pairs of focal papers have, and – as a consequence – in the number of female-authored focal papers included: all 1,261 papers in (A), 1,257 papers that could be paired based on at least one shared Faculty Opinions keyword in (B), 1,078 papers that could be paired based on at least two shared Faculty Opinions keywords in (C), 713 papers that could be paired based on at least three shared Faculty Opinions keywords in (D), 427 papers that could be paired based on at least four shared Faculty Opinions keywords in (E), and 210 papers that could be paired based on at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Share of female-authored citing papers instead of male-authored citing papers

In our main analyses, we focused on the share of male-authored citing papers in order to assess the degree of gender homophily in citation decisions. Since gender homophily in citations is the preference to cite authors of the same gender, gender homophily can also be operationalized by the difference in the share of female-authored citing papers. If female authors were more likely to cite other female authors, the share of female-authored citing papers would differ between male-authored and female-authored focal papers. Figure 2-21 shows the histograms for the differences in the share of female-authored citing papers for all pairs of focal papers from the Faculty Opinions data using Faculty Opinions keywords to control for paper similarity. Here, we calculated the differences as the share of female-authored citing papers for the male-

authored focal paper minus the share for the female-authored focal paper. This means that negative values indicate gender homophily in citations.

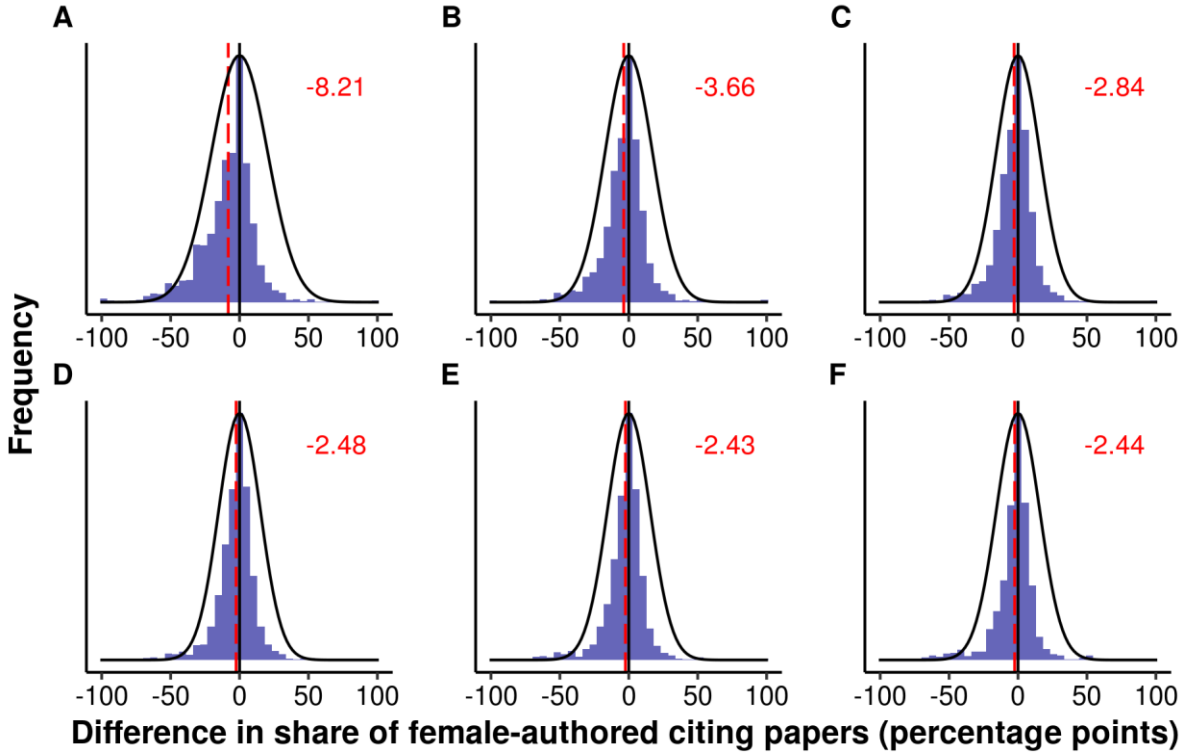


Figure 2-21. Histograms for the differences in the share of female-authored citing papers for pairs of focal papers (Faculty Opinions data).
In each histogram, the pairs of focal papers are restricted to those cases in which one focal paper is authored only by male scientists and the other focal paper is authored only by female scientists. Positive differences result when the female-authored paper of a pair has a higher share of female-authored citations than the male-authored paper of this pair. The histograms differ in the minimum number of shared keywords that the pairs of focal papers have, and – as a consequence – in the number of pairs of focal papers included: all 11,702,080 pairs in (A), 765,642 pairs with at least one shared Faculty Opinions keyword in (B), 223,837 pairs with at least two shared Faculty Opinions keywords in (C), 58,465 pairs with at least three shared Faculty Opinions keywords in (D), 14,167 pairs with at least four shared Faculty Opinions keywords in (E), and 3,010 pairs with at least five shared Faculty Opinions keywords in (F). The vertical lines are placed at 0 (black) and at the observed average difference (red, dashed). The black curve shows the shape of a normal distribution.

Without controlling for the similarity between papers, the difference between male-authored and female-authored focal papers in the share of female-authored citing papers is smaller than in the share of male-authored citing papers (8.21 vs. 12.64 percentage points). This may be due to the generally small share of female-authored citing papers: if both papers of a pair have a relatively small share of female-authored citing papers, the difference between them cannot be large either. In line with the analyses using the share of male-authored citing papers, the difference in the share of female-authored citing papers decreases when controlling for the similarity between papers. Most of the gender differences in citations disappears after controlling for the similarity between papers, and there is only a small degree of gender homophily in citations left once topic similarity is controlled. The remaining absolute difference is 2.4 percentage points,

which is close to the 1.6 percentage points when using male-authored citing papers (see Figure 2-4). The only major contrast to the share of male-authored citing papers is as follows: when using the share of female-authored (instead of male-authored) citing papers, already the first levels of topic similarity based on only one or two shared Faculty Opinions keywords are sufficient to net out nearly all gender differences. The difference hardly shrinks further when controlling for more than two Faculty Opinions keywords. One reason for this result may be the smaller average difference in the share of female-authored citing papers when not controlling for similarity: if there is only a small average difference, it cannot get much smaller any more when (further) controlling for the similarity.

These robustness checks (based on the difference in the share of female-authored citing papers) also support our main finding that controlling for the similarity between papers is important, even though the granularity of this similarity is not as important as when using the share of male-authored citing papers.

References

Araújo, E. B., Araújo, N. A. M., Moreira, A. A., Herrmann, H. J., & Andrade, J. S., Jr. (2017). Gender differences in scientific collaborations: Women are more egalitarian than men. *PLoS One*, *12*(5), e0176791. https://doi.org/10.1371/journal.pone.0176791

Backes, T. (2018). The impact of name-matching and blocking on author disambiguation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 803-812). ACM. https://doi.org/10.1145/3269206.3271699

Boekhout, H., van der Weijden, I., & Waltman, L. (2021). Gender differences in scientific careers. A large-scale bibliometric analysis. In W. Glänzel, S. Heeffer, P.-S. Chi, & R. Rousseau (Eds.), *Proceedings of the 18th International Conference on Scientometrics & Informetrics* (pp. 145-156). ISSI.

Bornmann, L. (2015). Interrater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for Information Science and Technology*, *66*(12), 2415-2426. https://doi.org/10.1002/asi.23334

Bornmann, L., Marx, W., & Barth, A. (2013). The normalization of citation counts based on classification systems. *Publications*, *1*(2), 78-86. https://doi.org/10.3390/publications1020078

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287-1294. https://doi.org/10.2307/1911963

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75-141. https://www.ncbi.nlm.nih.gov/pubmed/26172066

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, *108*(8), 3157-3162. https://doi.org/10.1073/pnas.1014871108

Chatterjee, P., & Werner, R. M. (2021). Gender disparity in citations in high-impact journal articles. *JAMA Netw Open*, *4*(7), e2114509. https://doi.org/10.1001/jamanetworkopen.2021.14509

Cook, R. D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. Taylor & Francis. https://books.google.de/books?id=MVSqAAAAIAAJ

Davenport, E., & Snyder, H. (1995). Who cites women? Whom do women cite?: An exploration of gender and scholarly citation in sociology. *Journal of Documentation*, *51*(4), 404-410. https://doi.org/10.1108/eb026958

de Kleijn, M., Jayabalasingham, B., Falk-Krzesinski, H. J., Collins, T., Kuiper-Hoyng, L., Cingolani, I., Zhang, J., Roberge, G., Deakin, G., Goodall, A., Whittington, K. B., Berghmans, S., Huggett, S., & Tobin, S. (2020). *The researcher journey through a gender lens: An examination of research participation, career progression and perceptions across the globe*. Retrieved 20 November 2022 from www.elsevier.com/gender-report

Dion, M. L., Sumner, J. L., & Mitchell, S. M. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, *26*(03), 312-327. https://doi.org/10.1017/pan.2018.12

Duch, J., Zeng, X. H., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., & Nunes Amaral, L. A. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One*, *7*(12), 1-11. https://doi.org/10.1371/journal.pone.0051332

Ferber, M. A. (1986). Citations: Are they an objective measure of scholarly merit? *Journal of Women in Culture and Society*, *11*(2), 381-389.

Ferber, M. A. (1988). Citations and networking. *Gender & Society*, *2*(1), 82-89.

Ferber, M. A., & Brün, M. (2011). The gender gap in citations: Does it persist? *Feminist Economics*, *17*(1), 151-158. https://doi.org/10.1080/13545701.2010.541857

Forscher, P. S., Cox, W. T. L., Brauer, M., & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature Human Behaviour*, *3*(3), 257-264. https://doi.org/10.1038/s41562-018-0517-y

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178-183.

Garfield, E. (1990). ISI's breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette. In *Essays of an Information Scientist: Journalology, KeyWords Plus, and other Essays* (Vol. 13, pp. 295-299). http://www.garfield.library.upenn.edu/essays/v13p295y1990.pdf

Garfield, E., & Sher, I. H. (1993). KeyWords Plus - algorithmic derivative indexing. *Journal of the American Society for Information Science*, *44*(5), 298-299.

Ghiasi, G., Mongeon, P., Sugimoto, C. R., & Larivière, V. (2018). *Gender homophily in citations.* Proceedings of the International Conference on Science and Technology Indicators (STI) 2018.

Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color? *Academic Medicine*, *91*(8). https://journals.lww.com/academicmedicine/Fulltext/2016/08000/Gender,_Race_Ethnicity,_and_National_Institutes_of.23.aspx

Glänzel, W., Thijs, B., & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, *59*(1), 63-77.

Håkanson, M. (2005). The impact of gender on citations: An analysis of College & Research Libraries, Journal of Academic Librarianship and Library Quarterly. *College & Research Libraries*, *66*(4), 312-323.

Halevi, G. (2019). Bibliometric studies on gender disparities in science. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 563-580). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_9

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*, 429-431. https://doi.org/10.1038/520429a

Holman, L., & Morandin, C. (2019). Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLoS One*, *14*(4), e0216128. https://doi.org/10.1371/journal.pone.0216128

Holman, L., Stuart-Fox, D., & Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented? *PLoS Biol*, *16*(4), e2004956. https://doi.org/10.1371/journal.pbio.2004956

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609-4616. https://doi.org/10.1073/pnas.1914221117

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, *21*(03n04), 1750011. https://doi.org/10.1142/s0219525917500114

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning. With Applications in R*. Springer.

Knobloch-Westerwick, S., & Glynn, C. J. (2013). The Matilda effect - role congruity effects on scholarly communication: A citation analysis of Communication Research and Journal of Communication articles. *Communication Research*, *40*(1), 3-26.

Larivière, V., Ni, C., Gingras, Y., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211-213. https://doi.org/10.1038/504211a

Lutz, C. (1990). The erasure of women's writing in sociocultural anthropology. *American Ethnologist*, *17*(4), 611-627.

Lynn, F. B., Noonan, M. C., Sauder, M., & Andersson, M. A. (2019). A rare case of gender parity in academia. *Social Forces*, *98*(2), 518-547.

Maliniak, D., Powers, R., & Walter, B. F. (2013). The gender citation gap in international relations. *International Organization*, *67*(4), 889-922. https://doi.org/10.1017/S0020818313000209

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*, *79*(3), 1290-1326. https://doi.org/10.3102/0034654309334143

Mcelhinny, B., Hols, M., Holtzkener, J., Unger, S., & Hicks, C. (2003). Gender, publication and citation in sociolinguistics and linguistic anthropology: The construction of a scholarly canon. *Language in Society*, *32*(2), 299-328.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415-444. https://doi.org/10.3410/f.725356294.793504070

Mitchell, S. M., Lange, S., & Brus, H. (2013). Gendered citation patterns in international relations journals. *International Studies Perspectives*, *14*(4), 485-492. https://doi.org/10.1111/insp.12026

Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Larivière, V. (2020). Who are the acknowledgees? An analysis of gender and academic status. *Quantitative Science Studies*, *1*(2), 582-598. https://doi.org/10.1162/qss_a_00036

Potthoff, M., & Zimmermann, F. (2017). Is there a gender-based fragmentation of communication science? An investigation of the reasons for the apparent gender homophily in citations. *Scientometrics*, *112*(2), 1047-1063. https://doi.org/10.1007/s11192-017-2392-0

Roper, R. L. (2019). Does gender bias still affect women in science? *Microbiology and molecular biology reviews : MMBR*, *83*(3), e00018-00019. https://doi.org/10.1128/MMBR.00018-19

Sammut, C., & Webb, G. I. (2010). TF–IDF. In *Encyclopedia of Machine Learning* (pp. 986-987). Springer US. https://doi.org/10.1007/978-0-387-30164-8_832

Stewart-Williams, S., & Halsey, L. G. (2021). Men, women and STEM: Why the differences and what should be done? *European Journal of Personality*, *35*(1), 3-39. https://doi.org/10.1177/0890207020962326

Studer, C. (2012). GitHub repository cstuder/genderReader. (March 20, 2019), Retrieved from https://github.com/cstuder/genderReader. https://github.com/cstuder/genderReader

Thelwall, M. (2018). Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries. *Journal of Informetrics*, *12*(4), 1031-1041. https://doi.org/https://doi.org/10.1016/j.joi.2018.08.005

Thelwall, M., Abdoli, M., Lebiedziewicz, A., & Bailey, C. (2020). Gender disparities in UK research publishing: Differences between fields, methods and topics. *El profesional de la información*, e290415. https://doi.org/10.3145/epi.2020.jul.15

Thijs, B. (2019). Science mapping and the identification of topics: Theoretical and methodological considerations. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 213-233). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_9

Traag, V., & Waltman, L. (2020). The causal intricacies of studying gender bias in science. *Leiden Madtrics*. Retrieved December 10 from https://leidenmadtrics.nl/articles/the-causal-intricacies-of-studying-gender-bias-in-science

van den Besselaar, P., & Sandström, U. (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLoS One*, *12*(8), e0183301. https://doi.org/10.1371/journal.pone.0183301

Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433-445. https://doi.org/https://doi.org/10.1002/asi.23040

Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 281-300). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_11

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS One*, *8*(7), e66212. https://www.ncbi.nlm.nih.gov/pubmed/23894278

Wildgaard, L. (2019). An overview of author-level indicators of research performance. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 361-396). Springer.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, *112*(17), 5360-5365. https://doi.org/10.1073/pnas.1418878112

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. Scientometrics, 126(11), 8861-8886. https://doi.org/10.1007/s11192-021-04171-y

3 Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example

Felix Bittmann, Alexander Tekles, Lutz Bornmann

Abstract

Controlling for confounding factors is one of the central aspects of quantitative research. Although methods such as linear regression models are common, their results can be misleading under certain conditions. We demonstrate how statistical matching can be utilized as an alternative that enables the inspection of post-matching balancing. This contribution serves as an empirical demonstration of matching in bibliometrics and discusses the advantages and potential pitfalls. We propose matching as an easy-to-use approach in bibliometrics to estimate effects and remove bias. To exemplify matching, we use data about papers published in *Physical Review E* and a selection classified as milestone papers. We analyze whether milestone papers score higher in terms of a proposed class of indicators for measuring disruptiveness than nonmilestone papers. We consider disruption indicators DI1, DI5, DI1n, DI5n, and DEP and test which of the disruption indicators performs best, based on the assumption that milestone papers should have higher disruption indicator values than nonmilestone papers. Four matching algorithms (propensity score matching (PSM), coarsened exact matching (CEM), entropy balancing (EB), and inverse probability weighting (IPTW)) are compared. We find that CEM and EB perform best regarding covariate balancing and DI5 and DEP performing well to evaluate disruptiveness of published papers.

3.1 Introduction

Scientometric research is mainly empirical research. Large-scale databases (e.g., Web of Science, Clarivate Analytics, or Scopus (Elsevier)) are used to investigate various phenomena in science. An overview of these studies can be found in (Fortunato et al., 2018). A popular topic of scientometric studies is the effect of gender. Researchers are interested in whether gender has an effect on the number of instances of being cited or the chance of being appointed for a professorship or fellowship. They want to know whether there is a systematic and robust gender bias in typical activities in science. Another popular topic of scientometric studies is the effect of the journal impact factor (a journal metric reflecting the reputation of a journal) on the citations of the papers published in a journal. Do papers profit from publication in a reputable journal in terms of being cited or not? An overview of studies that have investigated the relationship of journal impact factor and citations can be found in (Onodera & Yoshikane, 2015). Many of the studies investigating gender bias, citation advantages of the journal impact factor, and other phenomena have used multiple regression models to statistically analyze the data. In these models, the relationships between exactly one dependent variable (e.g., citation counts) and one or multiple independent variable(s) (e.g., journal impact factor) are investigated. Although in general regression methods are a valid tool to estimate (causal) effects, other methods can perform better in certain situations for multiple reasons, which will be outlined further below. In this paper, we present alternative methods—so-called matching techniques—which can be used instead of or as a supplement to regression models. It is our intention to explain the techniques based on a concrete empirical example for possible use in future scientometric studies.

Scientometric data are, as a rule, observational data (and not experimental data). Whenever observational data are available, simply comparing group means can create misleading results due to confounding influences. To achieve unbiased estimations of effects, various matching techniques exist to account for confounding. These techniques are usually referred to as *controlling* or *adjusting* to estimate unbiased effects balancing the distribution of covariates (possibly confounding factors) in the treatment and control groups (Paul, 1999; Rosenbaum, 2002; Rubin, 2007). Treatment groups are, for instance, female researchers/papers published by female researchers or papers published in reputable journals. Although statistical matching is not generally superior to methods such as regression models, and results can still be biased, if relevant confounders are omitted, they have several interesting properties that might be able to explain the growing popularity of matching techniques in various disciplines in recent years. These properties are outlined in detail in this study.

A few earlier studies by Farys and Wolbring (2017), Ginther and Heggeness (2020), Mutz and Daniel (2012), and Mutz et al. (2017) have demonstrated how useful matching techniques are for scientometric studies. For example, Mutz et al. (2017) have used the technique to investigate the effect of assigning the label "very important paper" to papers published in the journal *Angewandte Chemie—International Edition*. The authors were interested in whether this

assignment has a causal effect on the citation impact of the papers: Do these papers receive significantly more citations than comparable papers without this label? The results show that this is the case. In this study, we build upon these few previous studies and examine various matching techniques. Using a data set from bibliometrics as an exemplary case study, we explain various matching techniques in detail: propensity score-matching (PSM), inverse probability weighting (IPTW), coarsened-exact-matching (CEM), and entropy balancing (EB). The current paper can thus be understood as a methods paper explaining a certain statistic. In our opinion, the scientometric field would profit by applying these techniques more frequently in empirical research.

The example data that we used in this study are from *Physical Review E*—a journal focusing on collective phenomena of many-body systems. Editors of the journal denoted some papers from the journal as milestone papers in 2015. These milestone papers represent the treatment group in the current study. We are interested in whether this group of papers differs from a control group of papers in terms of indicators measuring disruptiveness of research. The goal of our analyses is to test how well the indicators perform: If the indicators adequately identify disruptive papers, the treatment and control group should differ with regard to the indicators. To compare the treatment group with a control group, four matching techniques are applied whereby several confounding variables are controlled in the statistical analyses, such as the number of coauthors of a paper and its number of cited references. The disruption indicators are recent developments in the field of scientometrics. By using the example data set with milestone papers from *Physical Review E*, the current study is a follow-up study of the study by Bornmann and Tekles (2021), who investigated milestone papers of the journal *Physical Review Letters* with the same set of indicators.

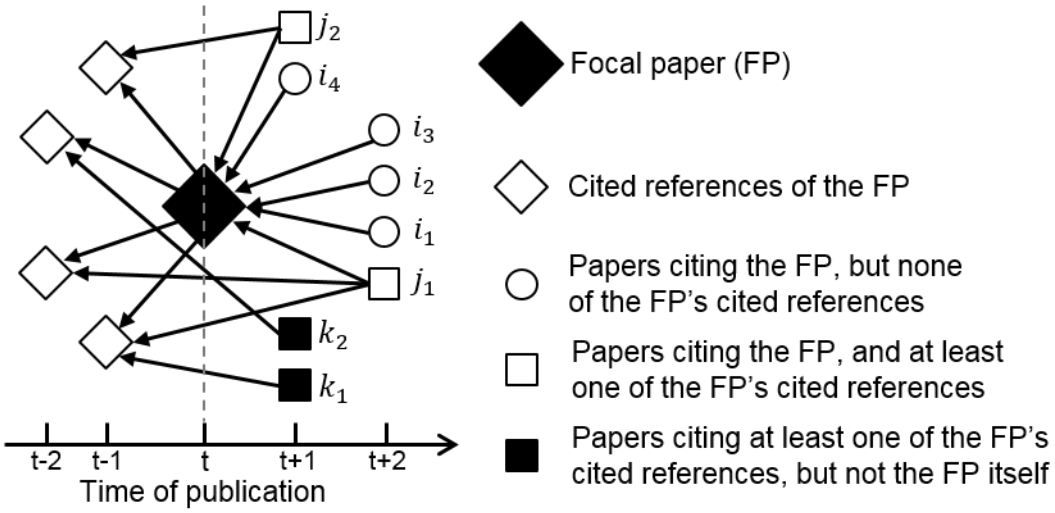
In the following sections, the example data set used in this study and the theoretical foundations of matching algorithms are described. Then, the matching results and results of the balancing and robustness checks are reported in the results section. In the last two sections of this paper, the matching procedures are finally discussed in the context of their application for bibliometric studies.

3.2 Dataset

For the generation of our data set, we started with a list of milestone papers published in *Physical Review E*. For this list, papers that made significant contributions to their field were selected by the editors of the journal. We assume that papers that made significant contributions to their field are more likely to be the origin of new lines of research (i.e., to be disruptive) than other papers. Based on this assumption, we want to test how well different indicators for measuring disruptiveness perform. To perform well, an indicator should on average differ between milestone and nonmilestone papers (we use the milestone assignment as a proxy for identifying papers that made significant contributions to their field). The papers in the list of *Physical Review E* milestone papers were published between 1993 and 2004. As this list was published in

2015, the selection of milestone papers may be influenced by citation information that was available by then. This possibility must be borne in mind when interpreting the results of our empirical analyses. To complete our data set for this study, we added all papers that are not in the list of milestone papers, but were also published in *Physical Review E* within the same time span. For all these papers, we retrieved additional bibliometric information from an in-house database at the Max Planck Society which is based on the Web of Science. For our analyses, we restricted the data set to the document type "article." This results in a list of 21,164 papers, of which 21 are milestone papers. Hence, the data set is very unbalanced with regard to the classification as milestone paper. Such data sets with a large difference in cases between treatment and control group are rather typical setups for the application of matching techniques. In clinical studies, for example, only a restricted number of ill or treated patients are available, with a large number of potential controls. These kinds of data sets are ideal for matching because the techniques make it possible to select the most appropriate controls out of a large pool of potential controls/donors. As others have pointed out, the control group should be larger than the treatment group by a factor of at least three, as this typically increases the common support region (in PSM) and allows finding multiple controls per treatment case (Olmos & Govindasamy, 2015, p. 86).

As we are interested in the difference between milestone and nonmilestone papers in terms of the indicators measuring disruptiveness, we used these indicators as outcome variables in our study. We considered five different indicators to measure the papers' disruptiveness: DI1, DI5, DI1n, DI5n, and the inverse DEP. These indicators all follow the same idea to measure disruptiveness: A focal paper (FP) can be regarded as disruptive if it is cited by many other papers that do *not* cite the FP's cited references. If this is the case, the citing papers depend on the FP but not its cited references (i.e., one can assume that the FP is the origin of new lines of research). In contrast, papers citing both the FP and its cited references indicate a developmental FP. This idea of measuring disruptiveness has been introduced recently in the context of patent analysis by Funk and Owen-Smith (2017). Wu et al. (2019) were the first to apply this concept to scientific publications by introducing the indicator DI1. The calculation of DI1 for a given FP is based on three terms (see Figure 3-1): N_i (the number of papers citing the FP but none of the FP's cited references), N_j^1 (the number of papers citing the FP and at least one of the FP's cited references) and N_k (the number of papers citing at least one of the FP's cited references but not the FP itself). The formula is based on the idea that N_i exceeds N_j^1 if the FP is disruptive. By including N_k, the indicator also considers how strong the citation impact of the FP is compared to its cited references.



$$DI_l = \frac{N_i - N_j^l}{N_i + N_j^l + N_k} \qquad DEP = \frac{N_{j\times cited}}{N_i + N_j^1}$$

$$N_i = |\{i_1, i_2, \ldots\}| \qquad N_j^l = |\{j_m | j_m \; cites \; FP \; and \; at \; least \; l \; of \; FP's \; cited \; references\}|$$

$$N_k = |\{k_1, k_2, \ldots\}| \qquad N_{j\times cited} = |\{(j_m, p_n) | p_n \; is \; cited \; by \; j_m \; and \; FP\}|$$

Figure 3-1. Definitions for disruption indexes DI1 and DI5 as well as the dependency indicator (DEP).

Since the introduction of DI1, several modifications of this indicator have been proposed. Out of these modified disruption indicators, we considered DI5, DI1n, and DI5n in this study because they showed good results in existing studies assessing their convergent validity (Bornmann et al., 2020a, 2020b; Bornmann & Tekles, 2021). In contrast to DI1, DI5 (which was first introduced in Bornmann et al., 2020a) considers how strong the ties between the citing and cited side of FPs are: A developmental FP is only indicated by citing papers that also cite at least five (instead of one) of the FP's cited references, which is captured in the term N_j^5 (see Figure 3-1). DI1n and DI5n are designed to measure the field-specific disruptiveness of a paper (Bornmann et al., 2020b). The definitions of DI1n and DI5n correspond to DI1 and DI5, respectively, but the FP's cited references are only considered for determining N_j^5 and N_k if they have been cited by other papers published in the same journal and the same year as the FP. All disruption indicators (DI1, DI5, DI1n, and DI5n) in their original form range from −1 to 1, with high (positive) values indicating disruptive papers (high negative values denote continuity in research). In this study, however, we multiplied the indicators by 100 for the statistical analyses to avoid small numbers with many decimal places. This transformation has been chosen to improve the presentation of the results.

Independently of the development of DI1, DI5, DI1n, and DI5n, Bu et al. (2021) proposed another indicator (DEP) that also follows the idea of considering whether the citing papers of

an FP cite the FP's cited references or not. Like DI5, DEP takes into account how strong the ties between the citing and the cited side of FPs are. More specifically, DEP is defined as the average number of citation links from a paper citing the FP to the FP's cited references (see Figure 3-1). A high (average) number of such citation links indicates a high dependency of citing papers on earlier work so that disruptiveness is represented by small values of DEP. In contrast to DI1, DI5, DI1n, and DI5n, DEP does not include a term for assessing the FP's citation impact (relative to the FP's cited references). This corresponds to a different notion of disruptiveness than DI1, DI5, DI1n, and DI5n build upon. DI1, DI5, DI1n, and DI5n follow the idea that FPs need to be relevant for a relatively large set of papers (compared to the FPs' cited references) in order to be disruptive. In contrast, the definition of DEP only considers to which extent citing papers refer to the cited references of FPs. To facilitate the comparison between DEP and DI1, and DI5, DI1n, and DI5n, we use the inverse DEP in this study, which is calculated by subtracting the values of DEP from the maximum value plus 1.

Since the introduction of the disruption indicators, some studies on their behavior and their validity have been published. Bornmann and Tekles (2019) have shown that it may take several years until the values of DI1 for a given paper reach a constant level. Therefore, a sufficiently long citation window is necessary to produce meaningful results (Bornmann & Tekles, 2019 suggest a citation window of at least three years). Because the data set of this study only comprises papers that were published in 2004 or earlier, this requirement is fulfilled in our statistical analyses. Other studies have shown that only very few papers score high on DI1, DI5, DI1n, and DI5n, whereas there are usually more papers with high values of the inverse DEP (Bornmann & Tekles, 2021).

Bornmann et al. (2020a) examined the convergent validity of the disruption indicators by analyzing the relationship between the indicator values and expert-based tags measuring newness of research. The study by Bornmann and Tekles (2021) used an external criterion for disruptive research similar to the current study to assess the convergent validity of the disruption indicators: a list of milestone papers published in the journal *Physical Review Letters* which were selected by the editors of the journal. Both of these studies found a considerable relationship between the disruption indicators and the external criteria for disruptiveness. However, both studies also found a stronger relationship between the external criteria for disruptiveness and citation impact. A similar finding was reported by Wei et al. (2020). The findings of these authors reveal that citation impact is a better predictor for Nobel prize-winning papers than disruptiveness in terms of DI1.

In the current study, we analyze whether milestone papers score higher in terms of the disruption indicators than the other papers published in the same journal. As the milestone papers were selected a few years after their publication, the citation impact may have played a role in the selection process. Therefore, the citation impact is very likely to be a good predictor for milestone papers. At the same time, the definitions of the disruption indicators also depend on citation patterns that may be related to citation impact and variables influencing the citation

impact. Thus, citation impact is a confounder for the effect of the milestone variable on the disruption indicators. To focus on this question, we compare the disruption indicator values of milestone and nonmilestone papers, which are comparable aside from the milestone assignment, by controlling the following variables in our analyses. These variables may have a considerable effect on citation impact.

The first variable is the number of coauthors. Due to the effects of self-citations and network effects (Valderas, 2007), this number might have an effect on citations, as different studies have demonstrated (e.g., Beaver, 2004; Fok & Franses, 2007; Tregenza, 2002; van Wesel et al., 2014) and thus be a potential confounder. In this study, we use the raw variable with values from 1 to 27. One extreme outlier from the control group with more than 100 coauthors is excluded.

The second control variable is the number of countries involved in a paper, which might have some effects regarding a national citation bias (Gingras & Khelfaoui, 2018). We transform this variable into a binary one (one country versus multiple countries) as there are only very few papers with many countries and it would be difficult to find appropriate matches.

The third variable is the age of each paper in terms of the years since publication. Older papers have had more time to be cited, which might influence their status (Seglen, 1992) and also the disruption indicator score (Bornmann & Tekles, 2019). This variable includes integers ranging from 1 to 12 years since publication.

The fourth control variable is the number of references cited by a paper. Multiple studies have shown a relation between the number of citations and the number of cited references (e.g., Ahlgren et al., 2018; Fok & Franses, 2007; Peters & van Raan, 1994; Yu & Yu, 2014). Although presumably not as relevant as in regular regression analyses, we use the log-transformed count of the number of references, as this gives a normally distributed variable which might be beneficial for the CEM cut-off algorithm.

Only papers with complete information on all relevant variables are retained for the statistical analyses (listwise deletion). Because the citation distributions of the milestone papers and the nonmilestone papers in our data set are very different, it is not possible to include the citation impact itself in the matching procedure. By restricting the data set to those papers that have at least as many citations as the least cited milestone paper, it is nevertheless possible to control for citation impact to a certain extent. We additionally used this restricted data set besides the data set including all papers to investigate the robustness of the empirical results.

3.3 Statistical matching

The general idea behind statistical matching is to simulate an experimental design when only observational data are available to make (causal) inferences. In an experiment, usually two groups are compared: treatment and control. The randomized allocation process in the experiment guarantees that both groups are similar, on average, with respect to observed and unobserved characteristics before the treatment is applied. Matching tries to mimic this process by

balancing known covariates in both groups. The balancing creates a statistical comparison where treatment and control are similar, at least with respect to measured covariates. If all relevant confounding factors are accounted for in statistical matching, causal effects can be estimated. Usually, balancing the observed covariates can help to balance unobserved covariates that are correlated with observed ones; hence, balancing is relevant for reaching high quality results (Caliendo & Kopeinig, 2008, p. 18). However, this cannot be proven statistically but must be defended with theoretical arguments. In the following, we present the advantages and challenges of statistical matching. We summarize various techniques that we empirically test using the example data set.

3.3.1 Advantages and disadvantages of statistical matching

Matching techniques have several advantages (compared to other statistics) for bibliometric analyses:

First, the techniques are conceptually close to the counterfactual framework (Morgan & Winship, 2014): Causal effects are estimated by generating a counterfactual situation whereby cases are observed with the nonfactual status (that is, treatment and control are swapped). In reality, however, this status does not exist. A case can only either have a treatment status or a control status. Matching approaches nevertheless follow this concept by comparing treated and untreated observations that are comparable with regard to the control variables considered. The idea behind the matching approach is that a treated (untreated) observation would, if it were untreated (treated), behave similarly to an actually untreated (treated) observation with comparable values for the control variables. This means for the empirical example of this study that a milestone paper would behave like a regular paper with similar values for certain control variables (number of coauthors, number of cited references, etc.). The only reason why the two papers behave differently is that one is a milestone paper and the other is not.

Second, the functional form of the relationship between treatment and outcome can be ignored. Although other methods such as linear regressions assume a strictly linear relationship and violations of this assumption can lead to severe biases in the results, matching is agnostic about this relation and reduces the number of specifications that the researcher has to check. This advantage is of special relevance for bibliometrics, as bibliometric data are usually concerned with skewed distributions.

Third, statistical matching allows the user to inspect the quality of the matching, which is an integral aspect of the validity of the estimated effects. Regression models can be considered to be rather opaque, as only regression coefficients are computed. Although the coefficients report the overall effect of a variable under the control of all other independent variables in the model, we are not informed about the validity of the findings. The computed coefficients might be based on highly dissimilar groups, which would invalidate the findings. With matching, the degree of similarity between treatment and control can be assessed after the procedure is performed. It can be examined whether the matching produced highly similar comparison groups

or not. If this assumption is violated in matching, the researcher knows that the results must be regarded with uttermost caution (the results probably cannot reveal any unbiased effects). For example, suppose that in a regression model a severe imbalance between treatment and control exists and, even after adjustment, a milestone paper has 10 authors on average and a regular paper only has two. The computed coefficient would be biased because this confounding factor could not be adjusted for. This is invisible to the user, however, who only sees the final coefficient and does not see how the groups were adjusted. Matching designs make these aspects transparent.

Fourth, in comparison to linear regressions that only report a single coefficient, matching allows the computation of multiple estimators with distinct meaning. Average treatment effects (ATEs) correspond to the regression coefficients (betas). ATEs can be interpreted as follows: Suppose a case is randomly selected for treatment. The effect is estimated as the counterfactual effect in comparison to the outcome that would have occurred if the case had been selected for the control group. In other words, the ATE is the effect for the "average" case in the sample. ATEs can be decomposed into ATT (average treatment effect on the treated) and ATC (average treatment effect on the control). ATT is the effect of treatment on those cases that actually received it, and ATC is the counterfactual effect of a case if it would have been treated. Hence, ATE is computed as the weighted mean of ATT and ATC. Depending on the research question, analyzing ATT, ATC, and their difference might be of special interest.

Like all other statistics, matching techniques have several disadvantages that should also be taken into account. The disadvantages are basically the counterparts to the advantages. As neither functional forms nor the separate contribution of control variables can be inspected, these techniques cannot replace regular regression designs. The techniques can be especially used for estimating treatment effects when the concrete functional form between treatment and outcome is irrelevant. Whenever a treatment is binary, this aspect can be ignored, as there is no functional form to be estimated. For other research questions dealing with *continuous* treatment variables, regression designs might be the better choice. In addition, regression techniques allow for the inspection of effects of multiple independent variables simultaneously, that is, under control of all other independent variables. This makes it possible to estimate how the independent variables *jointly* affect the outcome. In contrast, matching techniques only quantify the effect of the single treatment variable. All other control variables in the model are not further explained or quantified; coefficients are not computed for them. Furthermore, the functional form between treatment and outcome can be estimated using regression models. This functional form can be, for example, linear, quadratic, or exponential, depending on certain assumptions. The selection of the functional form is not possible for matching algorithms; they only compute single treatment effects. However, the functional form is often irrelevant in experimental designs, which matching algorithms attempt to mimic.

3.3.2 Matching for causal inference

Establishing causal relationships is one of the most important yet also most difficult aspects in data analysis, especially for policy-making and evaluation. Matching is a method that facilitates causal inference and especially causality according to Rubin (1974). In our case study, however, we are not interested in the analysis of a (potentially causal) effect of the milestone assignment on disruptiveness. Our goal is to test whether disruption indicators work as they are supposed to work. If this is the case, milestone assignments (a proxy for disruptiveness) should be associated with disruption indicator values. Therefore, matching approaches are a reasonable choice in this situation, because they allow us to control for the possible confounders mentioned in Section 3.2. By controlling confounders, associations between milestone assignments and disruption indicator values would not be due to confounding of control variables. Using matching approaches also allows us to assess matching quality. This is important in our case given the large control group (see also the advantages of matching approaches mentioned in Section 3.3.1).

With regard to using matching approaches for causal inference, we encourage the reader to have a look at the steadily growing body of literature and especially consult the works of Imbens and Rubin (2015), Morgan and Winship (2014), Pearl (2009), and Pearl et al. (2016). The authors target the social sciences and provide detailed examples. A nontechnical introduction for laypersons is given by Pearl and Mackenzie (2018). Whether or not the results of matching can be interpreted as causal effects depends on whether researchers are able to establish thoroughly that all assumptions for causal inference are indeed fulfilled. This can be achieved by theoretical and careful argumentation: No statistical test can derive whether or not a result is a causal effect. When researchers are not able to argue convincingly that all requirements are met, they should highlight the associational character of the findings. They cannot rule out hidden variable bias (for example).

3.3.3 An overview of various matching algorithms

After explaining the advantages and disadvantages of matching techniques in general, we present in the following an overview of various matching algorithms and explain their approaches to generate a balanced sample. Depending on the research questions, data sets, and designs of a certain bibliometric study, one of the matching algorithms might yield the most robust results. In this study, we apply four algorithms to the example data set; however, this is usually not feasible in a typical bibliometric paper. Our suggestion is therefore to compare at least a few algorithms with quite different statistical approaches (for example, CEM and EB) and inspect the quality of the findings. The selection of the algorithm should then be based on the most stable findings.

Propensity score matching (PSM)

To model the selection into treatment, a logistic (alternatively probit) model is used where the binary treatment status is the dependent variable and all potential confounders are independent

variables. The model computes the individual probability for each case to be selected for treatment as a number between 0 and 1 (Rosenbaum & Rubin, 1983). Because the potential confounders are relevant for the score, a case with a high individual propensity score has a high probability of being selected for treatment, even if the factual status is the control condition. Before matching, the region of common support for both treatment and control group should be reviewed: the computed propensity scores are compared between the groups. Only those cases are retained that have a value that is also available in the other group. For example, when the propensity score ranges from 5 to 60 in the control group and from 10 to 75 in the treatment group, the region of common support is from 10 to 60. There are no clear guidelines in the literature about whether imposing this restriction is always necessary, as it usually leads to a reduction of available cases. Modern implementations, in particular, of PSM, such as kernel-matching, usually do not benefit much from this restriction. In the analyses of this study we impose the common support restriction. After computing and restricting the propensity scores, cases are matched on it. For each case in the treatment group, one or multiple cases from the control group are selected, which should have an identical or very similar score.

Nearest-neighbor matching selects up to *n* neighbors for each treated case (Caliendo & Kopeinig, 2008). It is probably the most popular derivation of the general matching idea, as the assumptions are easy to comprehend, and it is implemented in many statistical software packages. By introducing a caliper (the maximum distance of two neighbors with respect to the propensity score), results can be improved as bad matches are avoided. By setting the caliper the user can adjust the balance between finding many matches and finding especially close matches. The mean differences in the outcome variable between matched cases can be compared to estimate the unbiased effect of the treatment. A similar propensity score guarantees that, on average, the cases are similar with regard to all control variables. More recent implementations rely on kernel instead of nearest-neighbor matching. Here, instead of selecting *n* neighbors, every single case is used but weighted by the degree of similarity (Bittmann, 2019). The closer the propensity score of a neighbor, the larger the weight. Although the introduction of kernel weighting usually improves the performance, reported case numbers can be deceptively large when many cases receive a weight close to zero (and contribute basically nothing to the estimation). Let us explain the technique based on our example data set. Instead of finding some similar control papers for a milestone paper which should be the nearest neighbors with respect to the propensity score, every single control paper is utilized as a neighbor. Then, only those control papers with a similar propensity score receive a high weighting, and other control papers with a highly different propensity score are discounted and receive a lower weighting. A very early implementation of the PSM approach is described in Rosenbaum and Rubin (1985). Further basic information on the approach can be found in Abadie and Imbens (2016), Heinrich et al. (2010), and Morgan and Winship (2014). If subgroup analyses are of interest in a study, these should be matched separately.

For the practical application of the technique, various software programs are available such as SPSS (Thoemmes, 2012), Stata (Jann, 2017a), and R (Olmos & Govindasamy, 2015; Randolph & Falbe, 2014). Although nowadays PSM is probably the most popular among the matching algorithms, some researchers argue that it might lead to an *increased* imbalance between groups (King & Nielsen, 2019) and might be inefficient (Frölich, 2007). Others counter that these downsides are only valid for rather crude PSM variants (one-to-one matching without replacement) and more recent implementations such as kernel matching do not display these problems (Jann, 2017b). In any case, due to its overall popularity and widespread use, we include PSM in this study and compare its performance with other algorithms. A further option to consider is the usage of regression adjustment, that is using the computed propensity score as a further control variable or stratifying the analyses based on propensity score levels (D'Agostino Jr., 1998).

Inverse probability weighting (IPTW)

Similar to PSM, IPTW relies on the propensity scores, which are calculated as described above; the same rules hold for selecting a region of common support. Each case receives a weight which is the inverse of the probability of receiving the factual status (Horvitz & Thompson, 1952). For example, case n_i in the treatment group receives the weight $w_i = 1/p_i(Treatment)$ whereby $p_i(Treatment)$ is the individual propensity score of this case. Cases in the control group receive the weighting $w_i = 1/[1 - p_i(Treatment)]$. That means that a case with a low probability of treatment in the treatment group receives a high weighting because it is similar to the untreated cases and enables a comparison. Cases with a high probability of treatment in the treatment group are weighted down, as there are many similar cases available with the same status. The calculation of the effect is then the weighted difference of means between the two groups. More information on the technique can be found in Austin and Stuart (2015) and Halpern (2014).

Coarsened exact matching (CEM)

Instead of relying on a propensity score, CEM attempts to find perfect matches. A perfect match occurs when there is a case available with a different treatment status but otherwise exactly the same characteristics (e.g., the same number of coauthors). Because the "curse of dimensionality" usually prevents the finding of perfect matches when the number of control variables is large, coarsening is used as a potential remedy (Iacus et al., 2012). For example, a continuous variable with a large number of distinct values is coarsened into a prespecified number of categories, such as quintiles. Matching is then performed based on quintile categories and the original information is retained. After matching based on the coarsened variables, the final effects are calculated as differences in the outcome variable between group means using the original and unchanged dependent variable.

The finer the degree of coarsening, the lower the number of potential matches. It is up to the user of CEM to test different coarsening rules and to find a balance between large numbers of

matches and high levels of detail and matching precision. For creating and selecting categories, multiple rules and algorithms are available. Suppose, for example, a user matches treatment and control papers based on their citation counts. As citation counts is a continuous variable, it might be impossible to find a perfect match for a paper with a specific number of citations, because no other paper in the control group has exactly this number. However, another paper is available having just one citation more. Through coarsening based on quintiles, both papers end up in the same quintile (a group of papers within a certain range of citation counts). The treatment paper with the specific number of citations has a match, therefore—albeit not a perfect match.

By coarsening, the aforementioned "curse of dimensionality" can be greatly ameliorated when many independent variables are included in a model. In our example data set, the binary variable "number of countries" is matched perfectly (because there are only two categories available and further coarsening is impossible). For more information on how to apply CEM, including practical examples, see Guarcello et al. (2017), Schurer et al. (2016), and Stevens et al. (2010).

Entropy balancing (EB)

In contrast to PSM, IPTW, and CEM, EB turns around the matching process. Instead of selecting similar cases and testing for balance afterwards, EB forces balancing with respect to prespecified conditions and generates matches according to the constraints by reweighting cases (Hainmueller, 2012). As this technique is highly flexible, the user can select various statistical moments that must be matched. These moments are usually means (first moment) and variances (second moment) of the independent variables. EB can be generalized to higher moments as well and some statistical packages allow matching of the skewness or even covariances.

After selecting the constraints, a loss function is used to meet the constraints. Each case receives a weight that is applied when group differences are computed. Constraints might not be met due to small sample sizes, a large number of constraints (matching multiple moments and covariances), or a strong imbalance between treatment and control group. If the constraints are not met, the algorithm does not converge and cannot yield an estimation. As a possible solution, the user can reduce the number of constraints. If the algorithm converges, the specified moments are basically guaranteed to be equal. The balancing should be close to an ideal state. A failure of balancing here might be a good indication for the user that other matching methods also provide suboptimal results. Further information on EB is available in Abadie et al. (2010), Amusa et al. (2019), and Zhao and Percival (2017).

3.3.4 Software

All the results presented in the following are computed using Stata 16.1 and the user-written software package *kmatch* (Jann, 2017a), which implements all of the matching algorithms described above. In the supplemental material, we also provide results computed using R as an additional robustness check (and to demonstrate that R can be equally used for matching as

Stata). For the R analyses, we used the R packages *MatchIt* (Ho et al., 2011), *ebal* (Hainmueller, 2014), and *boot* (Canty & Ripley, 2021).

3.4 Results

3.4.1 Descriptive statistics

Table 3-1 presents basic descriptive statistics for the milestone and regular papers included in this study. Although the asymmetry regarding the number of milestone papers to regular papers is extreme, the distribution of the control variables is very similar. For example, the number of coauthors involved and the number of cited references is comparable and not statistically significantly different between milestone and regular papers. Only the time since publication is statistically significantly different between both groups. In contrast to most of the control variables, most outcome variables display statistically significant differences between regular and milestone papers.

Figure 3-3 presents distributions of the outcome variables graphically using histograms. The histograms show that most of the values for DI1, DI5, DI1n, and DI5n lie in a small range around 0. There are only a few papers with relatively large or small values for these indicators. In contrast, the distribution for the inverse DEP indicator is less concentrated, even though most papers have values greater than 20. These results are in accord with the results of other empirical analyses concerned with disruption indicators (Bornmann & Tekles, 2021).

In addition, we use kernel-density plots to visualize how the citation counts differ between regular and milestone papers (see Figure 3-2).

The results in the figure reveal that milestone papers are cited more frequently than regular papers. Because we cannot include citation counts as a further control variable (see above), we run robustness checks where we remove all regular papers that have logarithmized citation counts below the lowest value of a milestone paper (5.69), which we indicate in the figure using a vertical bar.[2]

[2] Our initial analyses have shown that none of the matching algorithms is able to find an acceptable number of matches when this variable is included as independent variable. Therefore, we decided to use this approach.

Table 3-1. Descriptive statistics for the entire sample

	Minimum	Maximum	Mean	Standard deviation	Median
			Milestone papers (N = 21)		
Multiple countries involved	0.000	1.000	0.381	0.498	0.000
Number of co-authors	1.000	6.000	2.905	1.480	3.000
Years since publication (2005)	1.000	12.000	7.048*	3.263	7.000
Logarithmized number of cited references	2.565	4.331	3.516**	0.486	3.497
DI1 (DV)	-10.306	27.217	0.953	9.888	-2.826
DI5 (DV)	-0.663	32.702	7.333***	11.291	1.893
DI1n (DV)	-0.072	0.085	-0.023***	0.037	-0.030
DI5n (DV)	-0.028	0.125	0.015***	0.038	-0.001
DEP (inverse) (DV)	28.176	30.742	29.779**	0.815	29.962
			Regular papers (N = 21,143)		
Multiple countries involved	0.000	1.000	0.468	0.499	0.000
Number of co-authors	1.000	27.000	2.815	1.644	2.000
Years since publication (2005)	1.000	12.000	5.593	3.215	5.000
Logarithmized number of cited references	0.000	5.094	3.217	0.506	3.219
DI1 (DV)	-64.516	91.566	-0.636	2.676	-0.313
DI5 (DV)	-15.385	93.902	0.453	2.884	0.000
DI1n (DV)	-0.115	0.102	-0.001	0.003	-0.001
DI5n (DV)	-0.024	0.215	-0.000	0.002	-0.000
DEP (inverse) (DV)	1.000	31.000	28.325	2.127	28.765

Notes. Asterisks in column 'Mean' indicate whether group differences between regular and milestone papers are statistically significant (based on t-tests). Variables that are used as dependent variables in this study are marked with DV. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

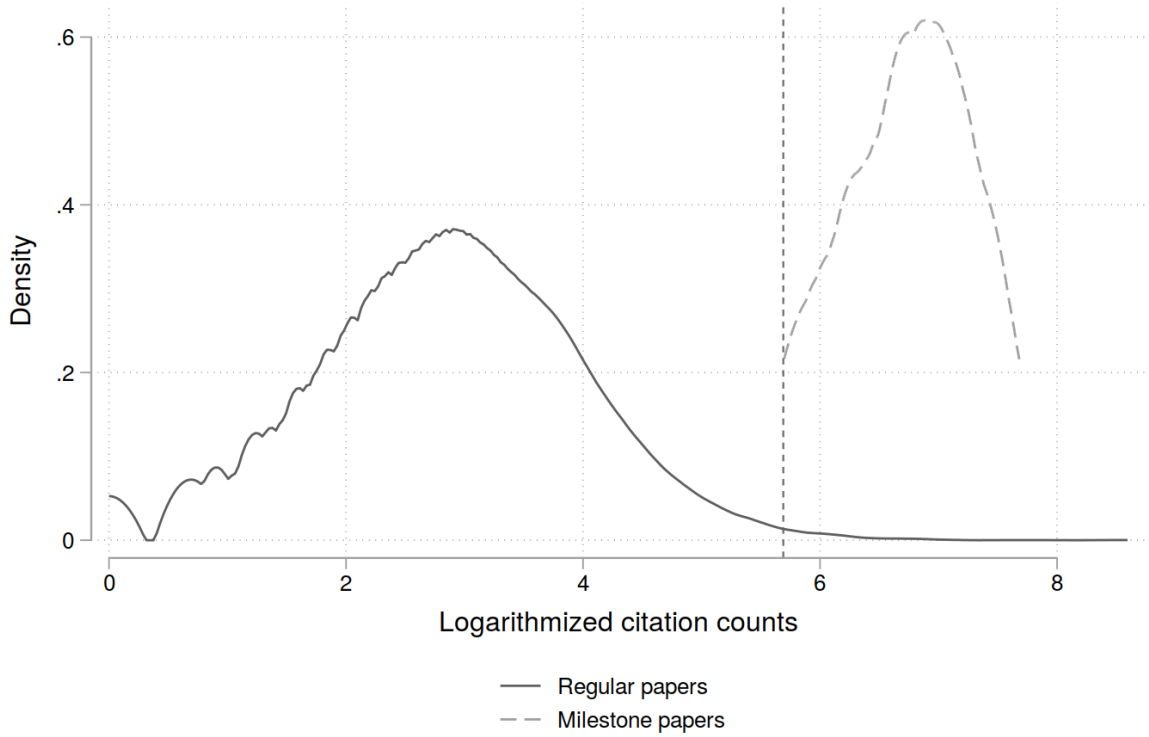


Figure 3-2. Comparison of regular and milestone papers with respect to logarithmized citation counts.
The lower limit of milestone papers is indicated by the vertical bar.

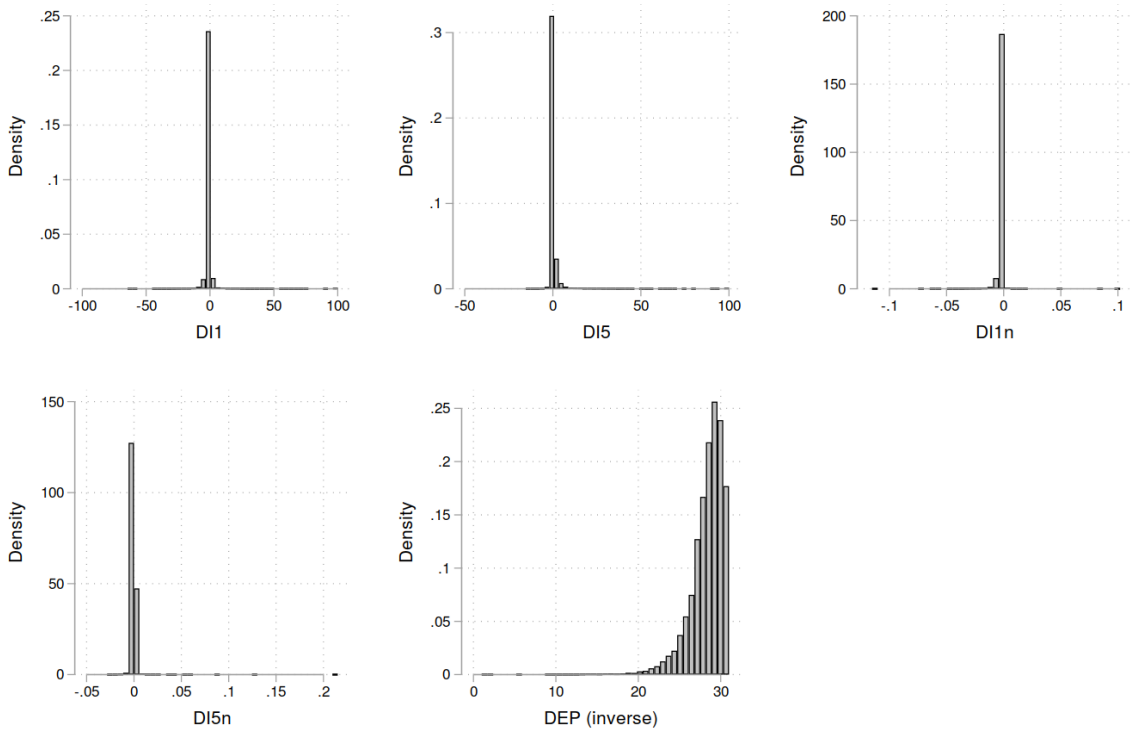


Figure 3-3. Distributions of all dependent variables.

3.4.2 Balancing and number of cases used

Before we discuss the treatment effect estimates of the various matching techniques based on the example data set in the following sections, we inspect the balancing, as this is relevant for judging the quality of the findings. All matching algorithms make it possible to inspect how well the observed covariates are balanced between treatment and control group. This is done by applying the computed weight to each case and recalculating the summary statistics of all independent variables. Balancing all control variables is a relevant aspect to obtain valid results. Even with balanced covariates, however, unmeasured variables might still be unbalanced and affect the validity of the estimation. When the balance of other influences (variables) is not approximated, a "fair" comparison between the groups is not possible as pretreatment differences are not completely accounted for. For a convenient interpretation of the balancing results, we create a single figure including all relevant information. We check the balancing for means, variances, and skewness. The means are the most relevant outcomes, as they are the first moment and determine the general shape of a distribution. The results are depicted in Figure 3-4.

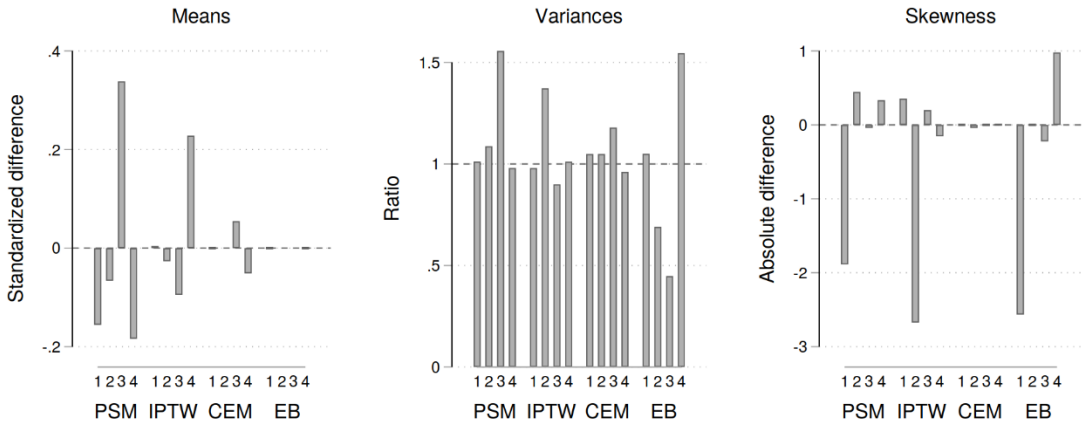


Figure 3-4. Inspecting balancing with respect to all independent variables between regular and milestone papers.
The IVs are enumerated from 1 to 4 (1=number of countries, 2=number of coauthors, 3=number of cited references, 4=number of years since publication).

The covariates are enumerated from 1 to 4 (1 = number of countries, 2 = number of coauthors, 3 = number of cited references, 4 = number of years since publication). "Means" reports the standardized difference between milestone and regular papers. When we look at the results for the PSM algorithm, we notice that differences regarding the means are quite large and can go up to 0.3. A perfect result would be close to zero. For the variances, the deviations are smaller

for most variables (a result of 1 would be ideal because we look at the ratios for this variable). For the third moment, the skewness, a few differences are large. We conclude that even after running the PSM models, some differences between the treatment and control groups remain. Perfect comparability with respect to all independent variables in the models cannot be guaranteed.

3.4.3 Results of the matching techniques

The actual matching outcomes of the four techniques are presented in Table 3-2. The table reports the average treatment effect (ATE) to measure the overall effect of treatment. Standard errors are computed analytically using influence functions (Jann, 2019). To test robustness, 95% confidence intervals are provided for the ATEs using bias-corrected bootstrapping with 2,000 resamples (Bittmann, 2021; Efron & Tibshirani, 1994). Because analytical standard errors can be too conservative for matching, we can test whether the conclusions are the same for both forms of computation (analytical and bootstrap standard errors) (Austin & Cafri, 2020; Hill, 2008; Jann, 2019). With ATEs, it is usually not possible to compute pure causal effects when only observational data are at hand and not all potential confounders are available. Therefore, the findings below can only be interpreted as rather associational than causal. To enable a comparison with the popular regression-based approaches, we also provide estimates for the treatment effects using ordinary least squares regression models in the supplemental material (Table 3-4).

Table 3-2 also reports the number of cases used in the statistical analysis. We notice that only CEM actually prunes many cases with a bad match (lower number of cases used). This is the only technique actually discounting controls, which are quite dissimilar with respect to the characteristics of their independent variables. All other techniques rely on some form of weighting and bad matches receive a very low weight. This means – as Table 3-2 reveals – that the estimated relationship between the milestone assignments and the indicator values is only based on very few papers with particular characteristics.

For the robustness check of our results (see Table 3-3), we compute the same models (including all matching techniques) but exclude all cases from the control group with rather low citation counts to enable a fair comparison (see above). This selection process drastically reduces the case numbers.

Table 3-2. Matching results

	PSM	IPTW	CEM	EB
DI1				
ATE	1.5786	1.4077	3.0627	1.9628
SE	(2.3082)	(1.9411)	(2.9518)	(2.1160)
95% analytical-CI	[-2.945; 6.102]	[-2.397; 5.212]	[-2.723; 8.848]	[-2.185; 6.110]
95% bootstrap-CI	[-2.787; 7.160]	[-2.762; 5.697]	[-0.861; 7.346]	-
DI5				
ATE	7.5612**	7.0657**	7.9726*	6.8608*
SE	(2.8426)	(2.5656)	(3.2481)	(3.2695)
95% analytical-CI	[1.989; 13.132]	[2.036; 12.094]	[1.606; 14.339]	[0.452; 13.269]
95% bootstrap-CI	[2.465; 14.362]	[2.585; 12.734]	[3.570; 13.159]	-
DI1n				
ATE	-0.0198**	-0.0197**	-0.0148	-0.0159*
SE	(0.0070)	(0.0065)	(0.0119)	(0.0069)
95% analytical-CI	[-0.033; -0.006]	[-0.032; -0.0069]	[-0.038; 0.009]	[-0.029; -0.002]
95% bootstrap-CI	[-0.033; 0.002]	[-0.0328; -0.0058]	[-0.030; 0.003]	-
DI5n				
ATE	0.0148*	0.0144*	0.0209	0.0135
SE	(0.0068)	(0.0068)	(0.0123)	(0.0077)
95% analytical-CI	[0.001; 0.028]	[0.001; 0.027]	[-0.003; 0.045]	[-0.001; 0.029]
95% bootstrap-CI	[0.002; 0.035]	[0.003; 0.030]	[0.006; 0.038]	-
DEP (inverse)				
ATE	1.7955***	1.7756***	1.7352***	1.6957***
SE	(0.1614)	(0.1509)	(0.2127)	(0.1987)
95% analytical-CI	[1.479; 2.111]	[1.479; 2.071]	[1.318; 2.152]	[1.306; 2.085]
95% bootstrap-CI	[1.315; 2.098]	[1.385; 2.204]	[1.437; 2.111]	-
Logarithmized citation counts				
ATE	3.7225***	3.6804***	3.7807***	3.6061***
SE	(0.1502)	(0.1276)	(0.1289)	(0.1339)
95% analytical-CI	[3.427; 4.016]	[3.430; 3.930]	[3.528; 4.033]	[3.345; 3.868]
95% bootstrap-CI	[3.464; 4.017]	[3.443; 3.952]	[3.537; 3.958]	-
N match (treated)	21	21	21	21
N match (control)	16,947	17,465	990	21,143

Notes. CI = Confidence Interval, ATE = Average Treatment Effect, SE = Standard Error, PSM = Propensity Score-Matching, CEM = Coarsened Exact Matching, EB = Entropy Balancing, IPTW = Inverse Probability Weighting. The outcome variables are the various disruption indicators (and citation counts) which can be found in the column on the left side. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3-3. Matching results (restricted sample)

	PSM	IPTW	CEM	EB
DI1				
ATE	3.5996	3.6531	4.2745	3.8060
SE	(2.3975)	(1.9176)	(6.1216)	(2.0458)
95% analytical-CI	[-1.135; 8.334]	[-0.133; 7.440]	[-7.815; 16.364]	[-0.234; 7.846]
95% bootstrap-CI	[-0.233; 9.488]	[0.118; 8.391]	-	-
DI5				
ATE	4.2480	4.6297*	7.3048	4.5821
SE	(2.4183)	(2.1567)	(6.3376)	(2.3608)
95% analytical-CI	[-0.527; 9.023]	[0.370; 8.888]	[-5.211; 19.821]	[-0.080; 9.244]
95% bootstrap-CI	[-0.095; 10.357]	[0.880; 9.915]	-	-
DI1n				
ATE	-0.0042	-0.0041	0.0222	-0.0018
SE	(0.0086)	(0.0073)	(0.0253)	(0.0076)
95% analytical-CI	[-0.021; 0.012]	[-0.018; 0.010]	[-0.0278; 0.072]	[-0.0169; 0.013]
95% bootstrap-CI	[-0.020; 0.017]	[-0.0179; 0.0136]	-	-
DI5n				
ATE	0.0139	0.0139	0.0526**	0.0127
SE	(0.0082)	(0.0072)	(0.0191)	(0.0078)
95% analytical-CI	[-0.002; 0.030]	[-0.001; 0.028]	[0.015; 0.090]	[-0.003; 0.028]
95% bootstrap-CI	[-0.001; 0.040]	[0.001; 0.032]	-	-
DEP (inverse)				
ATE	0.3430	0.3653	0.5410	0.4506*
SE	(0.1807)	(0.1871)	(0.3963)	(0.1791)
95% analytical-CI	[-0.013; 0.699]	[-0.004; 0.734]	[-0.242; 1.324]	[0.097; 0.804]
95% bootstrap-CI	[-0.2124; 0.689]	[-0.034; 0.704]	-	-
Logarithmized citation counts				
ATE	0.5472***	0.5410***	0.7542*	0.4887***
SE	(0.1344)	(0.1262)	(0.3694)	(0.1256)
95% analytical-CI	[0.2817; 0.812]	[0.291; 0.790]	[0.025; 1.484]	[0.241; 0.737]
95% bootstrap-CI	[0.256; 0.841]	[0.296; 0.825]	-	-
N match (treated)	20	21	4	21
N match (control)	131	133	5	140

Notes. CI = Confidence Interval, ATE = Average Treatment Effect, SE = Standard Error, PSM = Propensity Score-Matching, CEM = Coarsened Exact Matching, EB = Entropy Balancing, IPTW = Inverse Probability Weighting. Some confidence bands could not be calculated due to technical reasons. All regular papers with logarithmized citation counts below 5.69 are excluded. The outcome variables are the various disruption indicators (and citation counts) which can be found in the column on the left side. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Propensity score-matching (PSM)

We utilize a logistic model to compute the propensity score and kernel-matching to estimate ATEs. We restrict the region of common support and give a graphical representation of this process in the supplemental material (see Figure 3-5). This procedure is identical for PSM and IPTW. A review of the most popular software packages in Stata and R reveals that restricting the common support when applying kernel matching is not used as default and should only be imposed by the researcher if necessary.

The results in Table 3-2 show that PSM loses some cases due to restricting the common support region. The results reveal that five indicators have a statistically significant result when regular standard errors are computed: DI5, D15n, DI1n, the inverse DEP, and the logarithmized number of citations. The negative ATE of DI1n is an unexpected result, but probably not substantial. The bootstrap confidence interval (CI) does not agree, as zero is included in the interval. We cannot conclude, therefore, that a true relation is present. For the other statistically significantly independent variables the results of both CIs agree. According to the PSM technique, these three indicators should be rather robust. To test the stability of these findings, one option is to compute Rosenbaum bounds as a sensitivity analysis (Rosenbaum, 2002). The basic principle is to simulate the effect of unobserved variables on the selection into treatment and how this affects the results. If even a small additional effect of potential unobserved factors invalidates the findings (by changing *p*-values drastically), the results are probably not stable. We provide an exemplary analysis for the outcome variable DI1n in the supplemental material, see Table 3-5. When Gamma is 1, the *p*-value approximates the *p*-value of the average treatment effect from PSM as no unobserved influence is specified. The critical value of 0.05 is reached with a Gamma of 1.7: a change of 0.7 in odds in treatment assignment produces a statistically different result than the observed one. The larger the critical Gamma, the more robust the findings are with respect to unobserved influences that affect the treatment status (DiPrete & Gangl, 2004). As the value of 1.7 is not close to 1, we assume that the results are stable with respect to unobserved influences, even when the bootstrap CI is inconclusive.

The robustness check of the results (see Table 3-3) shows that only the citation count keeps its statistically significant result and DI5n is very close. Based on the results in Table 3-2 and their robustness checks in Table 3-3, we conclude that there are at most two disruption indicators with statistically significant results using the PSM technique. It should be considered in the interpretation of the results, however, that the balancing is not optimal.

Inverse probability weighting (IPTW)

With respect to the main findings in Table 3-2, we see that DI5, DI1n, DI5n, the inverse DEP, and citation counts display statistically significant coefficients. The table also shows a stable *negative* association for DI1n which is against our expectations. This means that milestone papers have a lower DI1n than the papers in the control group (on average). When we take a look

at the robustness checks in Table 3-3, this significance vanishes and only the findings of DI5 and citation counts remain stable.

The quality of the balancing indicates that the deviations for IPTW are the second largest after those for PSM. This result concerns the means but also the other moments. The balancing does not seem to be optimal even after the matching was performed. Thus, it appears that the IPTW results are not trustworthy.

Coarsened-exact-matching (CEM)

For control variables that are considered to be continuous, Doane's algorithm (1976) is selected to create categories. This formula does not simply generate equally spaced bins, but classifies cases into categories based on the distribution of the variable. Many other algorithms besides Doane's algorithm exist for this purpose. As no general standard has emerged hitherto, it is up to the user to try to compare various options for optimal results. We test different operationalizations and decide in favor of Doane's formula because the balancing of means and variances gives the best results.

The results in Table 3-2 show that DI5, the inverse DEP, and citation counts are statistically significant. For these variables, regular and bootstrap CIs agree, highlighting the stability of the results. The inspection of the robustness checks (see Table 3-3) reveals that DI5 is no longer statistically significant, DI5n becomes statistically significant, and citation counts remain statistically significant. Due to the very low number of cases used in the robustness checks (only nine in total), it is not feasible to compute bootstrap CIs. The regular results for the CEM are stable and robust; the robustness check might be neglected as the case number is very low.

The balancing in Figure 3-4 indicates that the deviations from the optimal results are very small for CEM. Treatment and control groups are very similar regarding the independent variables after the matching was performed. These results indicate the high quality of the matching process.

Entropy balancing (EB)

In our example data set, we select only the first moment (arithmetic means) as constraint as the model does not converge when we include higher moments as well. We assume that this is due to the very low number of milestone papers.

Table 3-2 indicates that the results for DI5, DI1n, the inverse DEP and citation counts are statistically significant. The negative association of DI1n is against the expectation. As the ATE is small and weak, it is probably not a robust finding. The coefficients for the inverse DEP and citation counts are large. The robustness checks in Table 3-3 show that only the inverse DEP and citation counts keep their statistical significances. The inspection of the balancing in Figure 3-4 reveals that deviations regarding the means are extremely small. This can be expected because the first moment can be matched very well. The figure also shows that the deviations for the variances and skewness are larger and worse than for CEM.

When we compare the results of EB with the results of CEM, we can conclude the following: Although the means are matched perfectly, the larger deviations with respect to variances and skewness speak for CEM. The balancing is better overall with CEM. Potential biases are probably smaller using CEM (i.e., the resulting conclusions are stronger using this algorithm).

3.5 Discussion

In this paper, we demonstrate how statistical matching techniques can be utilized as an addition or alternative to other methods, such as linear regressions. In contrast to these other methods, matching techniques are not only closer to the counterfactual framework but can sometimes be more adequate for analyses where the effect (or association in a noncausal framework) of exactly one (binary) treatment variable is of interest. Due to the different statistical approach, researchers not only estimate the desired statistic (in most cases the ATE) but are also able to study in detail how well treatment and control are matched after the procedure is performed. In contrast to linear regressions, where this aspect is opaque, researchers are able to conclude how well the matching was performed for the control variables specified and whether any larger bias is to be expected. By doing so, the quality of the results can be tested which is clearly highly relevant for scientific progress (in scientometrics). It is another advantage of statistical matching that the functional form between treatment and outcome can be ignored. We demonstrate in this study how matching can be applied in a practical example. We utilize bibliometric data to test which disruption indicator performs best. Several control variables are included to account for spurious correlations.

In this study, we use an example data set based on *Physical Review E* papers to demonstrate several matching techniques: PSM, IPTW, CEM, and EB. PSM and IPTW rely on the computation of the propensity score. This score is based on the control variables and predicts the propensity to be in the treatment group. CEM and EB have different requirements than PSM and IPTW. CEM implements an exact matching on broader categories. Depending on how the cut-off points for these categories are chosen, researchers are able to find a balance between precision and the number of cases left for analysis. EB attempts to force the balancing of covariates in advance. The balancing can fail to converge, however, if the number of cases is small and a good balancing solution is not feasible. If this happens, researchers can try to relax the balancing assumptions and can match only means and not variances.

In the empirical case study, we test with the matching techniques whether milestone papers differ from nonmilestone papers with regard to the various disruption indicators. Our results show that DI5, the inverse DEP and logarithmized citation counts have the strongest and robust results whereas outcomes for the other indicators are rather mixed. This suggests that these indicators perform best with regard to measuring the disruptiveness of papers. The found strong association for the number of citations is in line with the results from other studies (Bornmann et al., 2020a; Bornmann & Tekles, 2021). These results show that citation impact should ideally be controlled in the matching process to assess whether milestone and nonmilestone papers

differ with regard to the disruption indicators. In this study, this is not possible, as the citation distributions of the milestone and nonmilestone papers are very different in our data set.

Because citation counts themselves can not be included as control variable in the matching approaches, we performed robustness checks by restricting the papers to those with citation counts at least as high as the citation counts for the least cited milestone paper. This procedure makes it possible to control for citation impact to a certain degree. Among the disruption indicators, DI5 seems to perform best. This accords with existing studies that also find promising results for DI5 (Bornmann et al., 2020a; Bornmann & Tekles, 2021). In contrast to the existing studies, however, we also find promising results for the inverse DEP. The fact that DI5 and the inverse DEP perform best in our analyses may suggest that indicators measuring disruptiveness should take into account how strong the relationships between a citing paper and the cited references of a focal paper are, instead of only considering whether there is a citation link or not. Although DI5 and the inverse DEP both follow this idea, the approach to measure the field-specific disruptiveness of a paper (DI1n and DI5n) does not seem to be useful.

In this paper, we report results computed in R (in addition to Stata results) using the packages *MatchIt* (Ho et al., 2011), *ebal* (Hainmueller, 2014), and *boot* (Canty & Ripley, 2021); see supplemental material Table 3-6. The additional results reveal that the two software packages come to comparable results. The results underline that the implementations of matching techniques are equivalent and do not influence the conclusions.

In the application of the matching techniques in the empirical analyses, one is usually interested in which algorithm works best with the data. In this study, CEM and EB have the most robust and stable findings overall as well as the smallest deviations when looking at the balancing scores (see Figure 3-4). Here, deviations for the mean are small for both CEM and EB, which is the most relevant aspect when analyzing balancing statistics. Strong balance for derived statistics such as the variance and the skewness is also preferable but less relevant than balanced means. As both algorithms do rather well for all three measures (small deviations from a perfectly balanced sample after matching), we conclude that these two should be utilized as they minimize bias. In other words, both algorithms produce the best and most valid findings for our data set. PSM and IPTW display larger deviations regarding the matching of means for most variables. Even after the matching is performed, the difference of covariates between treatment and control group is comparably large. This can lead to biased and wrong conclusions. The propensity score might play a role in this context as this aspect is common to both. It is an advantage of matching techniques that we are able to test balancing and make this crucial aspect of the analyses visible (i.e., we can judge the final quality of the findings). This is not possible with many other techniques, such as linear regressions.

Previous studies compare the results of at least two techniques (Ginther & Heggeness, 2020) or combine different matching techniques (Farys & Wolbring, 2017). We would like to encourage researchers to follow these examples and we suggest applying more than one algorithm and comparing results. Modern software packages make it convenient to compute various

algorithms. Researchers strengthen the robustness of their results by doing so. Another option is to combine regression models with propensity score matching, which is referred to as a "double robust estimator" (Funk et al., 2011).

An important research gap that should be tackled in future studies is to compare algorithms systematically and to investigate how they perform with different scientometric data sets (e.g., with respect to the size of the treatment and control groups, total number of cases, and the number and kind of control variables used). Simulation studies might be helpful to find optimal algorithms for their analyses. The idea of such simulations is to generate a data set with known, prespecified effects which are set by the researcher. Then, the matching algorithms are applied to the data set to analyze whether they can recover the baseline truth. By repeating these simulations many times with varying conditions, the strengths and weaknesses of different algorithms can be tested systematically. As the number of potentially conceivable data sets is infinite, one would have to set very clear conditions. These conditions refer to the factors that should be evaluated and the specifications of the performance measurements for the algorithms. An example of using simulated data to validate a certain method in the field of scientometrics can be found in Milojević (2013), who applied this approach to assess different algorithms for identifying authors in bibliometric databases.

What are the limitations of our empirical analyses (the use of matching techniques in scientometrics)? Of course, it is not possible to report unbiased causal effects as only observational data are usually available in scientometrics (another assumption that it is necessary to inspect in a causal framework is strong ignorability). In addition, it is not possible to include every single confounding variable in a study that would be relevant in principle. In this study, for example, the citation impact can not be considered in the matching approaches. This is problematic because the disruption indicator values may be related to the number of citations that a paper receives and at the same time there is a strong relationship between citation impact and milestone paper assignments. Therefore, the estimated ATEs (without considering the citation impact as control variable) can be confounded by the citation impact. To still account for confounding by citation impact, we control for other variables that are related to citation impact. Another limitation concerns the extreme imbalance of the number of papers in the treatment and control groups. For the application of matching techniques, more balanced data sets should be used (ideally).

As in every scientometric study, empirical results must be interpreted with caution. The results of the current study give an estimation of how well the disruption indicators work. Whenever observational or nonexperimental data are available in scientometrics, it should be considered that causality cannot be proven statistically. Instead, one has to rely on theoretical reasoning for identifying relevant confounders (citation impact in our case). If one requires causal interpretation of the results, it is necessary to explain and outline plausibly that all potential confounding factors are accounted for. This necessity applies independently of the technique used and neither regular regression models nor matching are able to "prove" causality through statistics.

When the number of potential confounders is large in a scientometric study, it is possible to generate compound indices by using methods of data reduction. However, if central confounding factors are not available in the data or cannot be included in the matching process (such as citation impact in our case), one should refrain from causal interpretations. Thus, with respect to our data set, we are not able to interpret the computed statistics (ATEs) as unbiased causal effects but rather as associations.

3.6 Take-home messages

In this section, we summarize the most crucial aspects of using matching techniques in scientometric studies.

- Start by building your theoretical framework and formulate testable hypotheses. Name all potential confounders and describe how they are measured. When not all relevant confounders are measured, refrain from a causal interpretation of the results.
- Compute descriptive statistics for all dependent and independent variables you are going to use. The statistics help to choose the correct models and operationalization. For example, when mostly categorical covariates with few categories are used, exact matching might be a good solution. However, when the number of categories is large or continuous outcomes are utilized, tests of different algorithms to group these variables for CEM can be beneficial. Try to find a good balance between a reduction of bias and the number of cases left for analyses.
- Compute results for various matching techniques. This can be the most crucial aspect of the analyses because most techniques come with a large number of options. As these options also depend on the software package used to compute the results, inspecting the documentation is highly relevant. Either programmers themselves give recommendations for how to use certain options or you should test how strongly outcomes diverge when different options are utilized.
- Inspect balancing for each analysis. Report the results that minimize imbalance or report all results for comparison. Make sure to report balancing either using tables or graphics. If the balancing displays larger deviations between treatment and control group even after matching, the results might not be trustworthy and biases could be present. If the deviations are small, this does not mean that the results are unbiased (omitted variables could still have a confounding effect). It might show that balance is achieved between treatment and control group and there is no confounding left with respect to all control variables used in the models. This can be an iterative procedure: Insufficient balancing requires tweaking the matching model until sufficient balance is reached.
- Compute both regular (analytical) and bootstrapped standard errors for all coefficients of interest (e.g., ATE). The rationale behind this is to rely on two quite different assumptions. Regular standard errors are parametric and depend on the assumption of normality, which is rather strong. Bootstrapped standard errors require fewer

assumptions but more computational effort. When both standard errors come to similar conclusions, this points to the stability of the findings. If deviations between the standard errors are large, it should be checked whether there are underlying problems with the models or variables used. This might concern the skewness of continuous variables that deviates from the normal distribution. If no such obvious problems can be detected, report both types of standard errors and acknowledge that the results are potentially not very stable.

- When reporting the empirical findings, be transparent and describe the details of your results (software used, matching algorithms, imbalance, type of standard errors, etc.). Provide the source code (and raw data, if allowed) to aid replication studies.
- Use regression models as an additional robustness check. Regression techniques are highly popular for good reasons and reach beyond what matching can offer at the moment (for example, the consideration of various outcome variables and link families).

3.7 Appendix

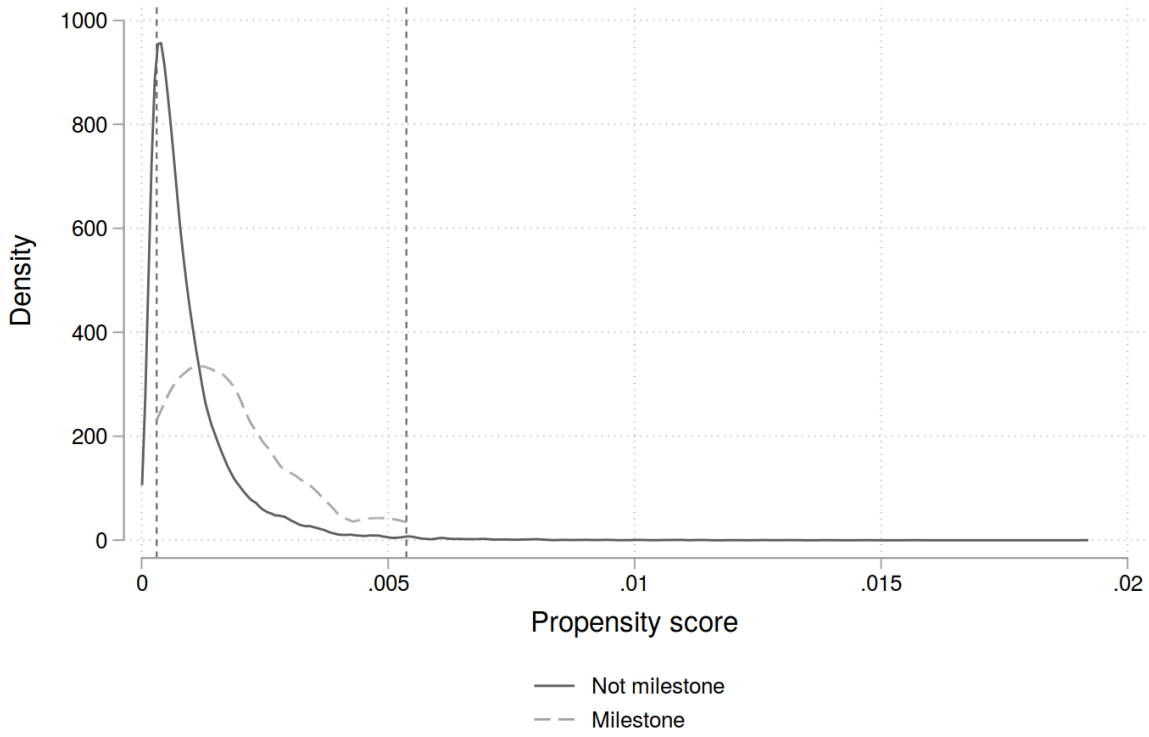


Figure 3-5. Distribution of generated propensity scores by milestone status.
The region of common support is the area between the two dashed lines as propensity scores completely overlap for both groups in this range.

Table 3-4. Treatment effects computed by ordinary least squares regression models

OLS	(1) DI1	(2) DI5	(3) DI1n	(4) DI5n	(5) Dep (inverse)	(6) Logarithmized citation counts
Milestone-paper	1.532	7.119***	-0.0214***	0.0153***	1.805***	3.661***
	(1.112)	(1.107)	(0.000594)	(0.000501)	(0.422)	(0.236)
N	22,082	22,082	22,086	22,086	21,164	22,086

Notes. Control variables are included but not depicted. Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3-5. Rosenbaum-bounds (p-values) computed for the outcome DI1n by PSM

Gamma (Γ)	P-value (lower)	P-value (upper)
1	0.004343	0.004343
1.1	0.002412	0.007429
1.2	0.001343	0.011645
1.3	0.000749	0.017059
1.4	0.000419	0.023692
1.5	0.000235	0.031524
1.6	0.000132	0.040503
1.7*	0.000074	0.050559
1.8	0.000042	0.061603
1.9	0.000023	0.073542
2.0	0.000013	0.086277

Notes. Findings computed in Stata using *rbounds*[3]. Critical gamma is marked with an asterisk as the p-value is above the reference bound (0.05). These results refer to 1:1 propensity score matching. Results might be unstable in comparison to kernel-matching. If the p-values computed for a gamma of 1 strongly deviate from the ones computed with kernel-matching, the sensitivity results are not adequate and should not be used.

[3] https://ideas.repec.org/c/boc/bocode/s438301.html

Table 3-6. Matching results using R

	PSM	IPTW	CEM	EB
DI1				
ATE	1.5782	1.4079	3.0507	1.9628
Bootstrap-SE	(1.9342)	(2.4193)	(2.5269)	(-)
95% Bootstrap-CI	[-2.784; 4.798]	[-3.295; 6.188]	[-2.108; 7.797]	-
DI5				
ATE	7.5608	7.0655	8.4055	6.8608
Bootstrap-SE	(2.9301)	(2.4729)	(3.1581)	(-)
95% Bootstrap-CI	[1.253; 12.739]	[1.713; 11.407]	[1.985; 14.364]	-
DI1n				
ATE	-0.0198	-0.0197	- 0.0111	-0.0159
Bootstrap-SE	(0.0067)	(0.0067)	(0.0086)	(-)
95% Bootstrap-CI	[-0.036; -0.010]	[-0.030; -0.004]	[-0.027; 0.007]	-
DI5n				
ATE	0.0148	0.0144	0.0228	0.0135
Bootstrap-SE	(0.0086)	(0.0059)	(0.0124)	(-)
95% Bootstrap-CI	[-0.004; 0.029]	[0.002; 0.026]	[0.000; 0.049]	-
DEP (inverse)				
ATE	1.7955	1.7752	1.7718	1.6957
Bootstrap-SE	(0.1789)	(0.1961)	(0.2263)	(-)
95% Bootstrap-CI	[1.338; 2.040]	[1.393; 2.162]	[1.254; 2.141]	-
Logarithmized citation counts				
ATE	3.7225	3.6806	3.7186	3.6061
Bootstrap-SE	(0.1464)	(0.1506)	(0.1476)	(-)
95% Bootstrap-CI	[3.485; 4.058]	[3.368; 3.959]	[3.479; 4.058]	-
N Match (Treated)	20	21	21	21
N Match (Control)	16929	17444	787	21143

Notes. CI = Confidence Interval, ATE = Average Treatment Effect, SE = Standard Error, PSM = Propensity Score-Matching, CEM = Coarsened Exact Matching, EB = Entropy Balancing, IPTW = Inverse Probability Weighting. The outcome variables are the various disruption indicators (and citation counts) which can be found in the column on the left side.

In addition to the results obtained using Stata and the *kmatch* package (which are presented in Table 3-2), we replicated the analyses using R. For this purpose, we used the R packages *MatchIt* (Ho et al., 2011) and *ebal* (Hainmueller, 2014). These packages allow to compute weights that can be used to calculate treatment effects. The packages do not support all matching approaches as described in the main text by default. Kernel matching (the particular PSM approach we applied) is not supported (we are not aware of any other R package that supports

kernel matching). Therefore, we implemented this approach based on propensity scores that can be calculated using the *MatchIt* package. When implementing this approach in R, we tried to use a setting as similar as possible to the *kmatch* package. In particular, we used the same kernel function and applied the same strategy to determine the bandwidth.

With regard to CEM, we used the default algorithm for determining the categories as provided by the MatchIt package, which is based on Sturges' formula (1926). In contrast, the *kmatch* package applies Doane's algorithm (which is not supported by the *MatchIt* package). In order to implement the IPTW approach, we used the propensity scores provided by the *MatchIt* package for calculating the inverse probability weights. The EB approach is implemented by the R package *ebal*, which also provides weights that can be used to calculate treatment effects. In align with our analyses using Stata, we consider only the first moment (arithmetic means) as balancing constraint. For the PSM and the IPTW approach, we restricted the sample to the region of common support.

In general, the treatment effects presented in Table 3-6 are very similar to those obtained using Stata. There are only minor differences between the Stata and the R implementation, suggesting that the results are robust with regard to different software packages. The largest difference can be observed for the CEM approach. This is probably due to the different strategies for determining the categories needed in the approach. Since the used R packages do not provide standard errors, we used the package *boot* (Canty & Ripley, 2021) to calculate bootstrapped standard errors and confidence intervals. Here, a few differences between the R implementation and the Stata implementation can be observed. Using the PSM approach, the confidence interval obtained using R does not include zero for DI1n, and includes zero for DI5n, while the opposite is true for the bootstrapped confidence intervals obtained using Stata. However, since the bootstrapped confidence intervals for DI1n do not accord with the analytical confidence intervals using Stata, these results already seem unreliable given only the Stata results.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493-505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781–807.

Ahlgren, P., Colliander, C., & Sjögårde, P. (2018). Exploring the relation between referencing practices and citation impact: A large-scale study based on Web of Science data. *Journal of the Association for Information Science and Technology*, *69*(5), 728-743. https://doi.org/10.1002/asi.23986

Amusa, L., Zewotir, T., & North, D. (2019). Examination of entropy balancing technique for estimating some standard measures of treatment effects: A simulation study. *Electronic Journal of Applied Statistical Analysis*, *12*(2). http://siba-ese.unisalento.it/index.php/ejasa/article/view/20409

Austin, P. C., & Cafri, G. (2020). Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in Medicine*, *39*(11), 1623-1640. https://doi.org/https://doi.org/10.1002/sim.8502

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661-3679. https://doi.org/https://doi.org/10.1002/sim.6607

Beaver, D. B. (2004). Does collaborative research have greater epistemic authority? *Scientometrics*, *60*(3), 399-408.

Bittmann, F. (2019). *Stata*. De Gruyter Oldenbourg. https://doi.org/doi:10.1515/9783110617160

Bittmann, F. (2021). *Bootstrapping*. De Gruyter. https://doi.org/doi:10.1515/9783110693348

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020a). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259. https://doi.org/10.1162/qss_a_00068

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020b). Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). *Scientometrics*, *123*(2), 1149-1155. https://doi.org/10.1007/s11192-020-03406-8

Bornmann, L., & Tekles, A. (2019). Disruption index depends on length of citation window. *Profesional de la información*, *28*(2). https://doi.org/10.3145/epi.2019.mar.07

Bornmann, L., & Tekles, A. (2021). Convergent validity of several indicators measuring disruptiveness with milestone assignments to physics papers by experts. *Journal of Informetrics*, *15*(3), 101159. https://doi.org/https://doi.org/10.1016/j.joi.2021.101159

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155-183. https://doi.org/10.1162/qss_a_00109

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72. https://doi.org/https://doi.org/10.1111/j.1467-6419.2007.00527.x

Canty, A., & Ripley, B. (2021). boot: Bootstrap R (S-Plus) functions. In R package version 1.3-28

D'Agostino Jr., R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*(19), 2265-2281. https://doi.org/https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34*(1), 271-310. https://doi.org/10.1111/j.0081-1750.2004.00154.x

Doane, D. P. (1976). Aesthetic frequency classifications. *The American Statistician*, *30*(4), 181-183. https://doi.org/10.1080/00031305.1976.10479172

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Farys, R., & Wolbring, T. (2017). Matched control groups for modeling events in citation data: An illustration of nobel prize effects in citation networks. *Journal of the Association for Information Science and Technology*, *68*(9), 2201-2210. https://doi.org/https://doi.org/10.1002/asi.23802

Fok, D., & Franses, P. H. (2007). Modeling the diffusion of scientific publications. *Journal of Econometrics*, *139*(2), 376-390, Article 1483. https://doi.org/10.1016/j.jeconom.2006.10.021

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Frölich, M. (2007). On the inefficiency of propensity score matching. *Advances in Statistical Analysis*, *91*(3), 279-290. https://doi.org/10.1007/s10182-007-0035-0

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761-767. https://doi.org/10.1093/aje/kwq439

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791-817. https://doi.org/10.1287/mnsc.2015.2366

Gingras, Y., & Khelfaoui, M. (2018). Assessing the effect of the United States' "citation advantage" on other countries' scientific impact as measured in the Web of Science (WoS) database. *Scientometrics*, *114*(2), 517-532. https://doi.org/10.1007/s11192-017-2593-6

Ginther, D. K., & Heggeness, M. L. (2020). Administrative discretion in scientific funding: Evidence from a prestigious postdoctoral training program. *Research Policy*, *49*(4), 103953. https://doi.org/https://doi.org/10.1016/j.respol.2020.103953

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental iInstruction through coarsened exact matching. *Technology, Knowledge and Learning*, *22*(3), 335-352. https://doi.org/10.1007/s10758-017-9317-0

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25-46. https://doi.org/10.1093/pan/mpr025

Hainmueller, J. (2014). ebal: Entropy reweighting to create balanced samples. R package version 0.1-6. https://CRAN.R-project.org/package=ebal

Halpern, E. F. (2014). Behind the numbers: Inverse probability weighting. *Radiology*, *271*(3), 625-628. https://doi.org/10.1148/radiol.14140035

Heinrich, C., Maffioli, A., & Vázquez, G. (2010). *A primer for applying propensity-score matching. Inter-American Development Bank*. https://publications.iadb.org/publications/english/document/A-Primer-for-Applying-Propensity-Score-Matching.pdf

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055-2061. https://doi.org/https://doi.org/10.1002/sim.3245

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1 - 28. https://doi.org/10.18637/jss.v042.i08

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663-685. https://doi.org/10.1080/01621459.1952.10483446

Iacus, S. M., King, G., & Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, *20*(1), 1-24. https://doi.org/10.1093/pan/mpr013

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9781139025751

Jann, B. (2017a). KMATCH: Stata module for multivariate-distance and propensity score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching and regression adjustment. In (Version S458346) Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458346.html

Jann, B. (2017b). *Why propensity scores should be used for matching* German Stata Users Group Meeting, Berlin.

Jann, B. (2019). Influence functions for linear regression (with an application to regression adjustment). https://doi.org/10.7892/boris.130362

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435-454. https://doi.org/10.1017/pan.2019.11

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, *7*(4), 767-773. https://doi.org/10.1016/j.joi.2013.06.006

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2 ed.). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9781107587991

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics*, *6*(2), 169-176. https://doi.org/https://doi.org/10.1016/j.joi.2011.12.006

Mutz, R., Wolbring, T., & Daniel, H.-D. (2017). The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology*, *68*(9), 2139-2153. https://doi.org/https://doi.org/10.1002/asi.23701

Olmos, A., & Govindasamy, P. (2015). Propensity scores: A practical introduction using R. *Journal of Multidisciplinary Evaluation*, *11*, 68-88.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, *66*(4), 739–764. https://doi.org/10.1002/asi.23209

Paul, R. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, *14*(3), 259-304. https://doi.org/10.1214/ss/1009212410

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96-146. https://doi.org/10.1214/09-SS057

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic books.

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science*, *45*(1), 39-49. https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<39::AID-ASI5>3.0.CO;2-Q

Randolph, J. J., & Falbe, K. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation*, *19*(18).

Rosenbaum, P. R. (2002). *Observational studies*. Springer. https://doi.org/https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38. https://doi.org/10.1080/00031305.1985.10479383

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701. https://doi.org/https://doi.org/10.1037/h0037350

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20-36. https://doi.org/https://doi.org/10.1002/sim.2739

Schurer, S., Alspach, M., MacRae, J., & Martin, G. (2016). The medical care costs of mood disorders: A coarsened exact matching approach. *Economic Record*, *92*(296), 81-93. https://doi.org/https://doi.org/10.1111/1475-4932.12218

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628-638. https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Stevens, G. A., King, G., & Shibuya, K. (2010). Deaths from heart failure: Using coarsened exact matching to correct cause-of-death statistics. *Population Health Metrics*, *8*(1), 1-9. https://doi.org/10.1186/1478-7954-8-6

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*(153), 65-66. https://doi.org/10.1080/01621459.1926.10502161

Thoemmes, F. (2012). *Propensity score matching in SPSS*. https://arxiv.org/pdf/1201.6385.pdf

Tregenza, T. (2002). Gender bias in the refereeing process? *Trends in Ecology & Evolution*, *17*(8), 349-350.

Valderas, J. M. (2007). Why do team-authored papers get cited more? *Science*, *317*(5844), 1496, Article 1492. https://doi.org/10.1126/science.317.5844.1496b

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: how superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601-1615. https://doi.org/10.1007/s11192-013-1154-x

Wei, C., Zhao, Z., Shi, D., & Li, J. (2020). *Nobel-prize-winning papers are significantly more highly-cited but not more disruptive than non-prize-winning counterparts*. https://www.ideals.illinois.edu/bitstream/handle/2142/106575/Contribution_477_final.pdf

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-382. https://doi.org/10.1038/s41586-019-0941-9

Yu, T., & Yu, G. (2014). Features of scientific papers and the relationships with their citation impact. *Malaysian Journal of Library & Information Science*, *19*(1), 37-50.

Zhao, Q., & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, *5*(1), 20160010. https://doi.org/10.1515/jci-2016-0010

4 Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches

Alexander Tekles, Lutz Bornmann

Abstract

Adequately disambiguating author names in bibliometric databases is a precondition for conducting reliable analyses at the author level. In the case of bibliometric studies that include many researchers, it is not possible to disambiguate each single researcher manually. Several approaches have been proposed for author name disambiguation, but there has not yet been a comparison of them under controlled conditions. In this study, we compare a set of unsupervised disambiguation approaches. Unsupervised approaches specify a model to assess the similarity of author mentions a priori instead of training a model with labeled data. To evaluate the approaches, we applied them to a set of author mentions annotated with a ResearcherID, this being an author identifier maintained by the researchers themselves. Apart from comparing the overall performance, we take a more detailed look at the role of the parametrization of the approaches and analyze the dependence of the results on the complexity of the disambiguation task. Furthermore, we examine which effects the differences in the set of metadata considered by the different approaches have on the disambiguation results. In the context of this study, the approach proposed by Caron and van Eck (2014) produced the best results.

4.1 Introduction

Bibliometric analyses of individuals require adequate authorship identification. For example, Clarivate Analytics annually publishes the names of highly cited researchers who have published the most papers belonging to the 1% most highly cited in their subject categories (see https://clarivate.com/webofsciencegroup/solutions/researcher-recognition/). The reliable attribution of papers to corresponding researchers is an absolute necessity for publishing this list of researchers. Empirical studies also showed that poorly disambiguated data may distort the results of analyses referring to the author level (Kim, 2019; Kim & Diesner, 2016). Some identifiers that uniquely represent authors are available in bibliometric databases. These are maintained by the researchers themselves (e.g., ResearcherID, ORCID) – implying a low coverage – or are based on an undisclosed automatic assignment (e.g., Scopus Author ID) – which does not allow an assessment of the quality of the algorithm (the algorithm is not publicly available). Publicly available approaches that try to solve the task of disambiguating author names have thus been proposed in bibliometrics. This task presents a nontrivial challenge, as different authors may have the same name (homonyms) and one author may publish under different names (synonyms).

Table 4-1 shows the titles, the author names and an author identifier for three publications, including both homonyms and synonyms. The author names of the first two publications are synonyms because they refer to the same person but differ in terms of the name. The author names of the last two publications are an example of homonyms because they refer to different persons but share the same name.

Table 4-1. Examples for homonyms and synonyms in bibliometric databases

Publication title	Author name	Author ID
Social theory and social structure	R. Merton	1
The Matthew effect in science	Robert Merton	1
Allocating Shareholder Capital to Pension Plans	Robert Merton	2

Although different disambiguation approaches have been developed and implemented in local bibliometric databases (e.g., Caron & van Eck, 2014), there is hardly any comparison of the approaches. However, this comparison is necessary to gain knowledge of which approaches perform best and the conditions on which the performance of the approaches depends. In this study, we compare four unsupervised disambiguation approaches. To evaluate the approaches, we applied them to a set of author mentions annotated with a ResearcherID, this being an author identifier maintained by the researchers themselves. Apart from comparing the overall performance, we take a more detailed look at the role of the parametrization of the approaches and analyze the dependence of the results on the complexity of the disambiguation task.

4.2 Related work

To find sets of publications corresponding to real-world authors, approaches for disambiguating author names try to assess the similarity between author mentions by exploiting metadata such as coauthors, subject categories, and journal. To reduce runtime complexity and exclude a high number of obvious false links between author mentions, most approaches reduce the search space by blocking the data in a first step (On et al., 2005). The idea is to generate disjunctive blocks so that author mentions in different blocks are very likely to refer to different identities, and therefore the comparisons can be limited to pairs of author mentions within the same block (Levin et al., 2012; Newcombe, 1967). A widely used blocking strategy for disambiguating author names in bibliometric databases is to group together all author mentions with an identical canonical representation of the author name, consisting of the first name initial and the surname (On et al., 2005; see also Section 4.4.1).

The algorithms to disambiguate author names that have been proposed up to now differ in several respects (Ferreira et al., 2012). One way to distinguish between different approaches is to classify them as either unsupervised or supervised (Smalheiser & Torvik, 2009). Supervised approaches try to train the parameters of a specified model with the help of certain training data (e.g., Ferreira et al., 2010; Ferreira et al., 2014; Levin et al., 2012; Torvik & Smalheiser, 2009). The training data contains explicit information as to which author mentions belong to the same identity and which do not. The model trained on the basis of this data is then used to detect relevant patterns in the rest of the data. Unsupervised approaches, on the other hand, try to assess the similarity of author mentions by explicitly specifying a similarity function based on the author mentions' attributes. Supervised approaches entail several problems, especially the challenge of providing adequate, reliable, and representative training data (Smalheiser & Torvik, 2009). Therefore, we focus on unsupervised approaches in the following.

The unsupervised approaches for disambiguating author names that have been proposed so far vary in several ways. First, every approach specifies a set of attributes and how these are combined to provide a similarity measure between author mentions. Second, to determine which similarities are high enough to consider two author mentions or two groups of author mentions as referring to the same author, some form of threshold for the similarity measure is necessary. This threshold can be determined globally for all pairs of author mentions being compared, or it can vary depending on the number of author mentions within a block that refer to a single name representation. Block-size-dependent thresholds try to reduce the problem of an increasing number of false links for a higher number of comparisons between author mentions; that is, for larger name blocks (Backes, 2018a; Caron & van Eck, 2014).

Third, the approaches differ with regard to the clustering strategy that is applied, that is, how similar author mentions are grouped together. All clustering strategies used so far in the context of author name disambiguation can be regarded as agglomerative clustering algorithms (Ferreira et al., 2012), especially in the form of single-link or average-link clustering. More

specifically, single-link approaches define the similarity of two clusters of author mentions as the maximum similarity of all pairs of author mentions belonging to the different clusters. The idea behind this technique is that each of an author's publications is similar to at least one of his or her other publications. In average-link approaches, on the other hand, the two clusters with the highest overall cohesion are merged in each step; that is, all objects in the clusters are considered (in contrast to just one from each cluster in single-link approaches). This rests on the assumption that an author's publications form a cohesive entity. As a consequence, it is easier to distinguish between two authors with slightly different oeuvres compared to single-link approaches, but heterogeneous oeuvres by a single author are more likely to be split.

Previous author name disambiguation approaches have usually been evaluated in terms of their quality. This evaluation is always based on measuring how pure the detected clusters are with respect to real-world authors (precision) and how well the author mentions of real-world authors are merged in the detected clusters (recall). However, different metrics have been applied when assessing these properties. Furthermore, different data sets have been used to evaluate author name disambiguation approaches (Kim, 2018). It is therefore difficult to compare the different approaches based on the existing evaluations.

4.3 Approaches compared

We focused on unsupervised disambiguation approaches in our analyses (see above). As these approaches require no training data to be provided a priori, they are more convenient for use with real-world applications. We investigated four elaborated approaches in addition to two naïve approaches, which only consider the author names (a) in the form of the canonical representation of author names used for the initial blocking of author mentions (first initial of the first name and the surname; see also Section 4.4.1), and (b) in the form of all first name initials and the surname. These approaches were selected to cover a wide variety of features that characterize unsupervised approaches for disambiguating author names. We applied the approaches to data from the Web of Science (WoS, Clarivate Analytics) that had already been preprocessed according to a blocking strategy, as described in Section 4.4.1.

4.3.1 Implementation of the four selected disambiguation approaches

In the following, the four disambiguation approaches that we investigated in this study are explained.

Cota, Gonçalves and Laender (2007) proposed a two-step approach that considers the names of coauthors, publication titles, and journal titles. In a first step, all pairs of author mentions that share a coauthor name are linked. The linked author mentions are then clustered by finding the connected components with regard to this matching. The second step iteratively merges these clusters if they are sufficiently similar with respect to their publication or journal titles. The similarity of two clusters (one for publication titles, one for journal titles) is defined as the cosine similarity of the two term frequency-inverse document frequencies (TF-IDFs) for the

clusters' publication titles (or journal titles). Two clusters are merged if one of their similarities (with either regard to publication or to journal titles) exceeds a predefined threshold. This process continues until there are no more sufficiently similar clusters to merge, or until all author mentions are merged into one cluster.

Schulz et al. (2014) proposed a three-step approach based on the following metric for the similarity s_{ij} between two author mentions i and j:

$$s_{ij} = \alpha_A \left(\frac{|A_i \cap A_j|}{\min(|A_i|,|A_j|)} \right) + \alpha_S \left(|p_i \cap R_j| + |p_j \cap R_i| \right) + \alpha_R \left(|R_i \cap R_j| \right) + \alpha_C \left(\frac{|C_i \cap C_j|}{\min(|C_i|,|C_j|)} \right) \quad \text{(I)}$$

Here, A_i denotes the coauthor list of paper i, R_i its reference list, and C_i its set of citing papers. The first step links all pairs of author mentions with a similarity (determined by Eq. (I)) exceeding a threshold β_1. A set of clusters is determined by finding the corresponding connected components. In the second step, these clusters are merged in a very similar way as in the first step. To determine the similarity $S_{\gamma\kappa}$ of two clusters γ and κ, the similarities between author mentions within these clusters are combined by means of the following formula:

$$S_{\gamma\kappa} = \sum_{i\in\gamma\, j\in\kappa} \frac{s_{ij}\Theta(s_{ij})}{|\gamma||\kappa|}, \quad \Theta(s_{ij}) = \begin{cases} 1 \; if \; s_{ij} > \beta_2 \\ 0 \; if \; s_{ij} \leq \beta_2 \end{cases} \quad \text{(II)}$$

Here, |γ| denotes the number of author mentions in cluster γ (similarly for cluster κ). As the formula shows, only those similarities between author mentions that exceed a threshold β_2 are considered when calculating the similarity between two clusters. As in the first step, this cluster similarity is used to link clusters if they exceed another threshold β_3 to find the corresponding connected components. The third step of this approach finally adds single author mentions that have not been merged to a cluster in either of the first two steps, provided its similarity with one of the cluster's author mentions exceeds a threshold β_4.

Caron and van Eck (2014) proposed measuring the similarity between two author mentions based on a set of rules that rely on several paper-level and author-level attributes. More precisely, a score is specified for each rule, and all of the scores for matching rules are added up to an overall similarity score for the two author mentions (see Table 4-2). If two author mentions are sufficiently similar with regard to this similarity score, they are linked and the corresponding connected components are considered oeuvres of real-world authors. The threshold for determining whether two author mentions are sufficiently similar depends on the size of the corresponding name block. The idea behind this approach is to take into account the higher risk of false links in larger blocks. Higher thresholds are therefore used for larger blocks to reduce the risk of incorrectly linked author mentions.

Table 4-2. Rules for rule-based scoring proposed by Caron and van Eck (2014)

Field	Criterion	Score
Email	exact match	100
Number of shared initials	2 / > 2 / conflicting initials	5 / 10 / -10
Shared first name	general name / non-general name	3 / 6
Address (linked to author)	matching country and city	4
Number of share co-authors	1 / 2 / > 2	4 / 7 / 10
Grant number	at least one shared grant number	10
Address (linked to publication, but not linked to author)	matching country and city	2
Subject category	matching subject category	3
Journal	matching journal	6
Self-citation	one publication citing the other	10
Bibliographic coupling: number of shared cited references	1 / 2 / 3 / 4 / > 4	2 / 4 / 6 / 8 / 10
Co-citation: number of shared citing papers	1 / 2 / 3 / 4 / > 4	2 / 3 / 4 / 5 / 6

Backes (2018a) proposed an approach that starts by considering each author mention as one cluster. An agglomerative clustering algorithm is then employed that iteratively merges clusters (starting with single author mentions as clusters, then merging clusters of several author mentions) if they are sufficiently similar; that is, two clusters are connected if their similarity exceeds a quality limit l. The similarity metric indicating how similar two clusters are takes into account the specificity of the author mentions' metadata. For example, if two author mentions share a very rare subject category this might be a strong indicator that the author mentions refer to the same author, while this is not true for a very common subject category. This strategy is applied to compute a similarity score for each attribute under consideration. The similarity score $p_a\left(C\middle|\dot{C}\right)$ for an attribute a and two clusters $C, \dot{C}$ is defined as

$$p_a\left(C\middle|\dot{C}\right) = \sum_{(x,\dot{x})\in C\times\dot{C}} p(x|\dot{x}) \cdot \frac{\#(\dot{x})+\varepsilon}{\#(\dot{C})+|C|\cdot\varepsilon} \quad \text{(III)}$$

with

$$p(x|\dot{x}) = \frac{1}{\#(\dot{x})+\varepsilon} \cdot \left(\left(\sum_{f\in F} \frac{\#(f,x)\cdot\#(f,\dot{x})}{\#(f)}\right) + \frac{\varepsilon}{|X|}\right),$$

$F =$ set of all features for attribute a,

$\#(f,x) =$ number of occurences of feature f for author mention x,

$\#(x) = \sum_{f\in F} \#(f,x)$,

$\#(C) = \sum_{x\in C} \#(x)$,

$\#(f) = \sum_{x \in X} \#(f, x)$,

$|X| =$ number of author mentions in the name block containing x and $\dot{x}$,

$\varepsilon =$ smoothing parameter to prevent division by zero.

When using this approach in our study, we considered the following attributes: titles, abstracts, affiliations, subject categories, keywords, coauthor names, author names of cited references, and email addresses. Backes (2018a) proposed several variants to combine these scores into a final similarity score of two clusters. In the variant implemented in this study, the scores are combined in the form of a linear combination with equal weights for all attributes' scores. This allows including attributes flexibly without the necessity to specify the corresponding weights separately. The results reported in Backes (2018a) suggest that using equal weights for all attributes produces good results. Each iteration of the clustering process merges all pairs of current clusters whose similarity exceeds l. The quality limit l is designed to have a linear dependence on the block size $|X|$, whereby the parameter λ specifies this relationship (see Eq. (IV)).

$$l = \lambda \cdot |X| \tag{IV}$$

Several other unsupervised approaches for disambiguating author names have been proposed besides the four aforementioned approaches (e.g., Hussain & Asghar, 2018; Liu et al., 2015; Wu et al., 2014; Wu & Ding, 2013; Zhu et al., 2017). Overviews of these approaches have been published by Ferreira et al. (2012) and Hussain and Asghar (2017). Our selection of the approaches aims at considering a wide range of strategies that can be applied for unsupervised author name disambiguation: using few versus using many attributes, using block-size-dependent versus using block-size-independent thresholds, and calculating similarity metrics based on various attributes versus merging author mentions based on one attribute at a time.

Besides the four approaches, we also included two naïve approaches that only use author names for the disambiguation. The first naïve approach uses the name blocks as the disambiguation result. This allows us to assess how much the elaborate approaches improve the disambiguation quality as compared to the blocking step alone. The second naïve approach only uses all initials of the first names and the surname for the disambiguation. This very simple approach has been widely used (Milojević, 2013) and seems to perform relatively well according to empirical analyses (Backes, 2018b). Including this approach in our analyses allows us to judge whether the additional effort associated with the more elaborate approaches is worthwhile with regard to the improvement in the disambiguation quality.

4.3.2 Parameter specification

Some form of threshold (or a set of thresholds) must be specified for each of the four approaches. As such thresholds have not been proposed for all approaches by the authors, and some of the proposed thresholds produce poor results for our data set, we fitted them with regard to our data. This allows better comparability because the thresholds are matched to the

particular data they are applied to. Our procedures for specifying the thresholds maximize the metrics $F1_{pair}$ and $F1_{best}$ (see below) that we used for the evaluation of the approaches. In our analyses, this is primarily a means for evaluating the approaches independently of the particular thresholds used, as the results reflect how good the approaches are instead of how well the thresholds are chosen. In practical applications, this would only be possible if a sufficiently large amount of the data is already reliably disambiguated (which is usually not the case though).

We specified a procedure for each of the approaches that allowed an efficient consideration of a wide range of thresholds. A set of thresholds uniformly distributed over the complete parameter space was chosen as a candidate set for the approach of Cota, Gonçalves and Laender (2007). We also specified the thresholds for the approach of Schulz et al. (2014) by evaluating a candidate set of parameters; in this case, the candidate set of thresholds was chosen on the basis of the parameters proposed in the original paper. The parametrization of this approach was further optimized by fitting β_1, β_2, and β_3 independently from β_4. β_4 was subsequently chosen based only on the best combination of the other thresholds, which substantially reduces the search space. We believe this to be an adequate procedure for finding the thresholds because the last step of this disambiguation approach (which is based on β_4) has only a minor influence on the final result. For the approach proposed by Caron and van Eck (2014) we initially had to define the block size classes that divide the blocks into several classes with regard to the internal number of author mentions. Similar to Caron and van Eck (2014), we defined six block size classes. Our specification of the classes aims at reducing the variance of optimal thresholds within a class and is based on a manual inspection of the distribution of optimal thresholds across block sizes. Then the best possible threshold for each class (maximizing $F1_{pair}$ and $F1_{best}$) is chosen.

For the approach of Backes (2018a), we had to modify the approach slightly to define a feasible procedure for fitting the parameter λ, which determines the quality limit l for a given block. Instead of linking all pairs of clusters whose similarity exceeds a given l in each iteration, we iteratively merged only those pairs of clusters whose similarity equals the maximum similarity of all current pairs of clusters (the clusters are recomputed after each merger). These similarities were taken as estimates for the quality limit that would yield the clustering of the corresponding merger step. This modification may produce results that are different to the original approach, because the order in which the author mentions are merged may change and the similarities between clusters depend on the previous mergers. However, we assume that these changes produce only minor differences that do not influence any general conclusions on the approach. Our implementation merges the most similar clusters in each iteration; that is, the most reliable mergers are applied iteratively until the quality limit is reached. Correspondingly, the original approach follows the idea that all cluster similarities exceeding a certain quality limit indicate reliable links between the corresponding clusters.

4.4 Method

We collected metadata for a subset of author mentions from the WoS for our analyses. To provide a gold standard that represents sets of author mentions corresponding to real-world authors, we only took author mentions with a ResearcherID linked to their publications in the WoS into account. More specifically, all person records that are marked as authors and that have a ResearcherID linked to at least one paper published in 2015 or later have been considered. It is very likely that this procedure excludes all author mentions with ResearcherIDs referring to nonauthor entities (e.g., organizations) and takes into account only such ResearcherIDs that have been maintained recently.

For an increasing number of author mentions, it can be expected that the quality of their disambiguation decreases (see also Section 4.5). Our results would thus not be transferable to application scenarios with a larger number of author mentions than in our data set. At the same time, the limitation on a subset of author mentions from the WoS seems appropriate, because the same data is used for all approaches. This allows comparing the approaches under controlled conditions. Furthermore, our analyses allow an assessment of the relationship between the complexity of the disambiguation task (in terms of name block size) and the quality of the results produced (see Section 4.5). This gives an idea of how well the approaches perform for an increasing amount of data. As including more author mentions in our data would drastically increase the computational costs, we refrained from including more author mentions than those annotated with a ResearcherID.

4.4.1 Blocking

Blocking author mentions based on authors' names is usually the first step in the disambiguation process. While different strategies have been proposed for this blocking step, they all aim at narrowing down the search space for the subsequent disambiguation task in a reliable and efficient way. For this purpose, a canonical representation of the author name is specified and all author mentions with identical name representation are assigned to the same block.

As this procedure only considers author names and is based on exact matches, it requires less computational resources compared to the subsequent steps of the disambiguation process. These subsequent steps can be applied then to smaller sets of author mentions. Because the computational complexity of the disambiguation approaches considered in our study is superlinear in the number of author mentions, the overall complexity can be reduced by splitting up the disambiguation in smaller tasks. A smaller number of author mentions also reduces the risk of making false links between author mentions, which improves the quality of the disambiguation results.

While reducing the block sizes, the blocking strategy at the same time needs to be reliable in the sense that for an author, a canonical name representation is very likely to include all of her or his author mentions. To achieve both goals, an adequate level of specificity of the canonical name representation used for blocking the author mentions is necessary. Using a general name

representation (e.g., the first initial of the first names and the full surname) results in relatively large blocks. The number of splitting errors is rather small in these blocks, but the computational complexity of the subsequent steps in the disambiguation process is rather high. In contrast, using a specific name representation (e.g., all initials of the first names and the full surname) results in smaller blocks. Although the number of splitting errors in these blocks increases due to synonyms, the computational complexity of the subsequent steps is reduced in the disambiguation process. Empirical analyses assessing the errors introduced by different blocking schemes can be found in Backes (2018b). These analyses show that a general name representation based on the first initial of the first names and the full surname produces good results, especially with regard to recall. They also show that using all initials of the first names and the full surname produces good results in terms of F1 (see Section 4.4.2). These results qualify the blocking scheme based on all initials of the first names and the full surname as a simple disambiguation approach without any subsequent steps. However, compared to using only the first initial and the surname, blocking the author mentions based on all initials of the first names and the full surname introduces additional splitting errors. These splitting errors introduced by the blocking step are of particular importance for subsequent steps, because they cannot be corrected later in the disambiguation process.

For the blocking step in our analyses, we used the first initial of the first names and the full surname as the canonical name representation. One reason for this choice is that this name representation has been used by many other studies related to author name disambiguation (Milojević, 2013). A second reason is that this is a very general blocking scheme, which reduces the risk of making splitting errors in the blocking step (Backes, 2018b). For a practical application with a large amount of data, this might not be feasible, because the general blocking scheme produces large blocks (Backes, 2018b). However, for our purpose of evaluating different approaches building upon the blocked author mentions, using a general blocking scheme allows us to focus on these subsequent steps. Due to the high recall, the upper bound for the disambiguation quality that can be achieved by the approaches is not reduced considerably by the blocking step, and the final result is more dependent on the subsequent steps rather than the blocking step. The small risk of making splitting errors due to this blocking scheme is also visible in our results (see Table 4-3).

In our analyses, we only considered name blocks comprising at least five real-world authors. This selection allowed us to focus on rather difficult cases where the author mentions in a block actually have to be disambiguated across several authors. All in all, this data collection procedure results in 1,057,978 author mentions distributed over 2,484 name blocks and 29,244 distinct ResearcherIDs. The largest name block (“y. wang”) comprises 7,296 author mentions.

4.4.2 Evaluation metrics

The evaluation of author name disambiguation approaches is generally based on assessing their ability to discriminate between author mentions of different real-world authors (precision) and

their ability to merge author mentions of the same real-world author (recall). Even though these concepts are widely accepted and referenced, various specific evaluation metrics have been used in the past. In the following, we focus on two types of evaluation metrics. First, we calculate pairwise precision (P_{pair}), pairwise recall (R_{pair}), and pairwise F1 ($F1_{pair}$) for each approach. These metrics have been used in many studies (e.g., Backes, 2018a; Caron & van Eck, 2014; Levin et al., 2012).

Whereas pairwise precision measures how many links between author mentions in detected clusters are correct, pairwise recall measures how many links between author mentions of real-world authors are correctly detected. Pairwise F1 is the harmonic mean of these two metrics. Eqs. (V)-(VII) provide a formal definition of these evaluation metrics, using the following notation:

- $|pairs_{author}|$ denotes the number of all pairs of author mentions where both author mentions refer to the same author
- $|pairs_{cluster}|$ denotes the number of pairs of author mentions where both author mentions are assigned to the same cluster by the disambiguation algorithm
- $|pairs_{author} \cap pairs_{cluster}|$ denotes the number of author mentions where both author mentions refer to the same author and are assigned to the same cluster

$$P_{pair} = \frac{|pairs_{author} \cap pairs_{cluster}|}{|pairs_{cluster}|} \quad \text{(V)}$$

$$R_{pair} = \frac{|pairs_{author} \cap pairs_{cluster}|}{|pairs_{author}|} \quad \text{(VI)}$$

$$F1_{pair} = \frac{2P_{pair}R_{pair}}{P_{pair} + R_{pair}} \quad \text{(VII)}$$

An important property of pairwise evaluation metrics is that they consider the disambiguation quality among all links between author mentions. For example, consider two clusters A and B for which the precision should be determined. Cluster A has 10 author mentions referring to one author and five author mentions to a second author. Cluster B has 10 author mentions referring to one author and five author mentions referring to different authors. These two clusters get different scores for the pairwise precision (for cluster A, $P_{pair} = \frac{55}{105} \approx 0.524$, while for cluster B, $P_{pair} = \frac{45}{105} \approx 0.429$). However, if we assign each cluster to one author, the two clusters are equally adequate: Ten author mentions are correct and five are incorrect in each case. To assess how the disambiguation approaches perform with regard to this task (and the corresponding task to find all author mentions for each author), we calculated metrics to measure how reliably a cluster can be attributed to exactly one author (best precision P_{best}) and how well an author can be attributed to exactly one cluster (best recall R_{best}). Eqs. (VIII)-(X) provide a formal definition of these evaluation metrics, using the following notation:

- $|author\ mentions_{best\ author}|$ is calculated as follows: for each cluster c, the maximum number of author mentions that refer to the same author $n_{c,\max author}$ is determined; $|author\ mentions_{best\ author}|$ is the sum of $n_{c,\max author}$ over all clusters
- $|author\ mentions_{best\ cluster}|$ is calculated as follows: for each author a, the maximum number $n_{a,\max cluser}$ of author mentions that are assigned to the same cluster is determined; $|author\ mentions_{best\ cluster}|$ is the sum of $n_{a,\max cluser}$ over all authors
- $|author\ mentions|$ denotes the number of all author mentions

$$P_{best} = \frac{|author\ mentions_{best\ author}|}{|author\ mentions|} \quad \text{(VIII)}$$

$$R_{best} = \frac{|author\ mentions_{best\ cluster}|}{|author\ mentions|} \quad \text{(IX)}$$

$$F1_{best} = \frac{2P_{best}R_{best}}{P_{best} + R_{best}} \quad \text{(X)}$$

An approach for evaluating the quality of author name disambiguation that is very similar to P_{best}, R_{best}, and $F1_{best}$ has been proposed by Li et al. (2014). In this approach, splitting and lumping errors are calculated, which correspond to the notions recall and precision, respectively. However, the calculation of lumping errors does not necessarily take into account all clusters, but for each author the cluster with most of her or his author mentions. In contrast, P_{best} considers all clusters. Therefore, P_{best} is better suited to assess how reliable it is to take each cluster as one author given the disambiguated data (see also Torvik & Smalheiser, 2009 for a discussion of different perspectives for evaluating author name disambiguation). Furthermore, P_{best}, R_{best}, and $F1_{best}$ are better comparable with the pairwise evaluation metrics, because both types of metrics follow the precision-recall-F1 terminology and have the same scale. Another type of evaluation metrics that are very similar to P_{best}, R_{best}, and $F1_{best}$ are the closest cluster precision, closest cluster recall, and closest cluster F1 (Menestrina et al., 2010). These metrics are based on the Jaccard similarities between clusters and authors.[4] The closest cluster precision is the average maximum Jaccard similarity over all clusters. By using the maximum Jaccard similarities for each cluster, this approach is very similar to the idea that P_{best} is based on: For each cluster, only the author with the most author mentions in this cluster is taken into account[5]. However, in contrast to P_{best}, a closest cluster precision < 1 is possible if each cluster only contains author mentions of one author. When considering such a cluster as the oeuvre of one author, the precision should be 1 though: All author mentions in this cluster are correct (all author mentions refer to the same author, that is, the cluster is perfectly precise).

[4] The Jaccard similarity $J(a, c)$ between author a and cluster c is defined as $\frac{number\ of\ author\ mentions\ in\ c\ and\ a}{number\ of\ author\ mentions\ in\ c\ or\ a}$

[5] The closest cluster recall is calculated accordingly by changing the perspective from clusters to authors.

Therefore, we decided to use P_{best}, R_{best}, and $F1_{best}$ as defined in Eqs. (VIII)-(X) for evaluating the disambiguation approaches in this study.

Each of Eqs. (V)-(X) can be applied either to the complete data set or to a subset of author mentions. For example, the results of one name block can be evaluated by only considering author mentions within this block when computing the evaluation metrics. All metrics can take values between 0 and 1, with higher values indicating a better disambiguation result.

4.5 Results

4.5.1 Overall results

The results for the approaches described in Section 4.3 are summarized in Table 4-3. The table shows the evaluation metrics described in the previous section for each approach. All the approaches produced better results than the naïve baseline disambiguation based on first initial and surname; only three of the approaches produced better results than the baseline disambiguation based on all initials and surname. The approach proposed by Caron and van Eck (2014) performs best among the examined approaches with regard to both $F1_{pair}$ and $F1_{best}$. If one compares the approaches of Schulz et al. (2014) and Backes (2018a), the two evaluation metrics yield different rankings. Whereas the latter approach performs better with regard to $F1_{pair}$, the former performs better with regard to $F1_{best}$. Both of these approaches perform only slightly better than the baseline based on all initials. This might suggest that a simple approach based only on author names performs nearly as well as these approaches. However, the precision of the all-initials baseline is very small compared to the approaches of Schulz et al. (2014) and Backes (2018a). The all-initials baseline and the two approaches also differ in the variance of the disambiguation quality across block sizes (see Figure 4-1). This means that the approaches perform better or worse depending on the given data and the preferences regarding the trade-off between precision and recall. The approach of Cota et al. (2007) performs worse than the all-initials baseline, and only slightly better than the first-initial baseline. The precision in particular is very small for the approach of Cota et al. (2007), mainly due to a high number of false links between author mentions in the first step (merging author mentions with shared coauthors).

Table 4-3. Overall results for all approaches

Approach	P_{pair}	R_{pair}	$F1_{pair}$	P_{best}	R_{best}	$F1_{best}$
Baseline (first initial)	0.095	0.998	0.173	0.322	0.999	0.487
Baseline (all initials)	0.210	0.854	0.338	0.603	0.905	0.724
Cota et al. (2007)	0.111	0.857	0.196	0.442	0.912	0.595
Schulz et al. (2014)	0.453	0.456	0.455	0.799	0.749	0.773
Caron and van Eck (2014)	0.831	0.785	0.808	0.916	0.884	0.900
Backes (2018a)	0.674	0.620	0.646	0.761	0.698	0.728

Figure 4-1 shows the distribution of the disambiguation quality over block sizes, using thresholds as described in Section 4.3.2. The lines represent nonparametric regression estimates (calculated using the `loess()` function in the base package of R), with evaluation metrics as dependent variable and block size as independent variable. In addition to these regression estimates, the results for single blocks are plotted for large block sizes. As there are too many small blocks to adequately recognize the relationship between block length and evaluation metrics, results at the block level are only displayed for large blocks.

The results reveal that the disambiguation quality in terms of the F1 metrics varies strongly across name blocks. In particular, the F1 values decrease for large blocks. Therefore, the disambiguation process may produce biases with regard to the frequency of the corresponding name representation. One reason for the dependence of the disambiguation quality on the size of the name block is the larger search space to find clusters of author mentions. The larger search space increases the search complexity in general, implying a greater potential for false links between author mentions. Some approaches try to reduce this problem by allowing block size-dependent thresholds (see the next section). Even though the negative relationship between block size and disambiguation quality can be observed for all approaches, the decline in quality is not equal. Especially for the approach of Caron and van Eck (2014), the influence of the block size is relatively small.

Besides the scores for the F1 metrics, Figure 4-1 also shows the distribution of (pairwise) precision and recall values across block sizes. According to these results, the approach of Caron and van Eck (2014) favors precision over recall, even for large blocks. The approach of Backes (2018a) scores very high on the precision metrics, but very low on the recall metrics for large blocks. This suggests that the specification of thresholds only works for small blocks in this case (see the next section). The other approaches produce results with rather small precision for large blocks, while their recall values are relatively high.

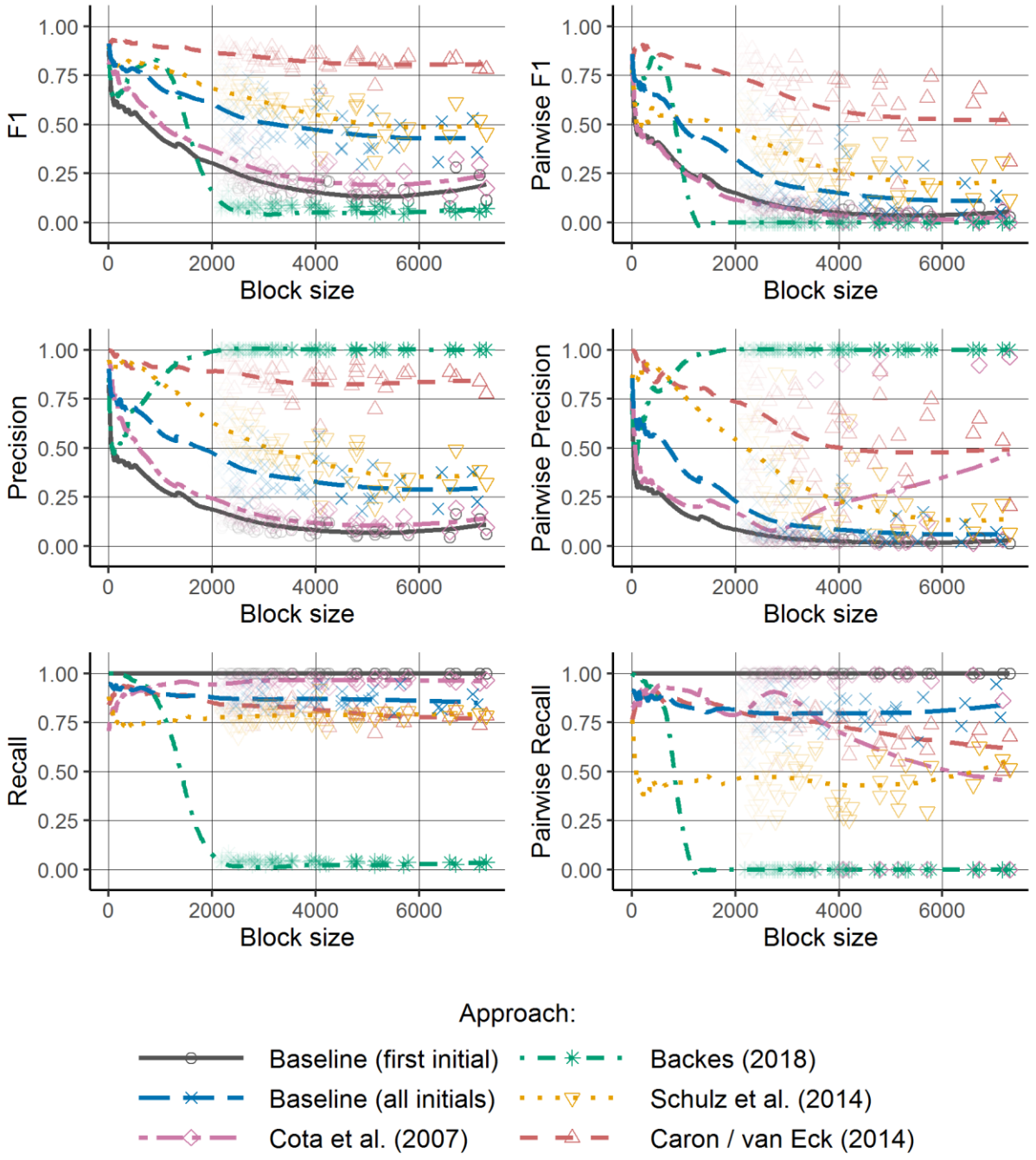


Figure 4-1. F1, precision and recall values for all approaches across block sizes using thresholds as originally proposed by the authors.
The lines show nonparametric regression estimates based on all blocks and the points show results for single blocks (only results for large blocks are displayed this way).

4.5.2 The influence of parametrization on the disambiguation quality

Among the approaches included in our comparison, Caron and van Eck (2014) and Backes (2018a) used block-size-dependent thresholds. As described above, the first approach is based on defining one threshold for each of six block size classes, whereas the threshold is linearly dependent on the block size in the second approach. Table 4-4 shows the block size classes and corresponding thresholds used by our implementation for the approach of Caron and van Eck

(2014). In contrast, the approaches of both Cota et al. (2007) and Schulz et al. (2014) use global thresholds for all block sizes.

Table 4-4. Block size classes and thresholds for Caron and van Eck (2014)

Block size	Threshold ($F1_{pair}$)	Threshold ($F1_{best}$)
1-500	21	19
501-1000	22	21
1001-2000	25	23
2001-3000	27	25
3001-4500	29	25
>4500	29	27

To assess how much the results could be improved by allowing different thresholds for the blocks, we determined the thresholds producing the best result for each block. Figure 4-2 shows the evaluation results obtained by using these optimal thresholds for each single name block – instead of using the same thresholds for (a) all blocks, (b) a group of blocks, or (c) determining the thresholds based on a global rule as described in section 4.3.2. These results represent an upper bound for the quality over all possible thresholds if the thresholds are specified for each name block separately. The difference in the results between Figure 4-1 (using thresholds as originally proposed) and Figure 4-2 (using flexible thresholds) indicates the improvement potential for each approach by optimizing how the thresholds are specified. As the specification of flexible thresholds requires reliably disambiguated data beforehand, this strategy is not feasible in application scenarios. Flexible thresholds for each block would not greatly improve the quality of the approach proposed by Cota et al. (2007) because the results based on global thresholds are very close to the results based on completely flexible thresholds. The reason is that the quality is dominated by the first step of the approach, which does not employ any threshold at all. The second step, on the other hand, does not change the results significantly; the effect of the thresholds is rather small. In contrast, the approach of Schulz et al. (2014) benefits from using flexible thresholds, especially for large blocks.

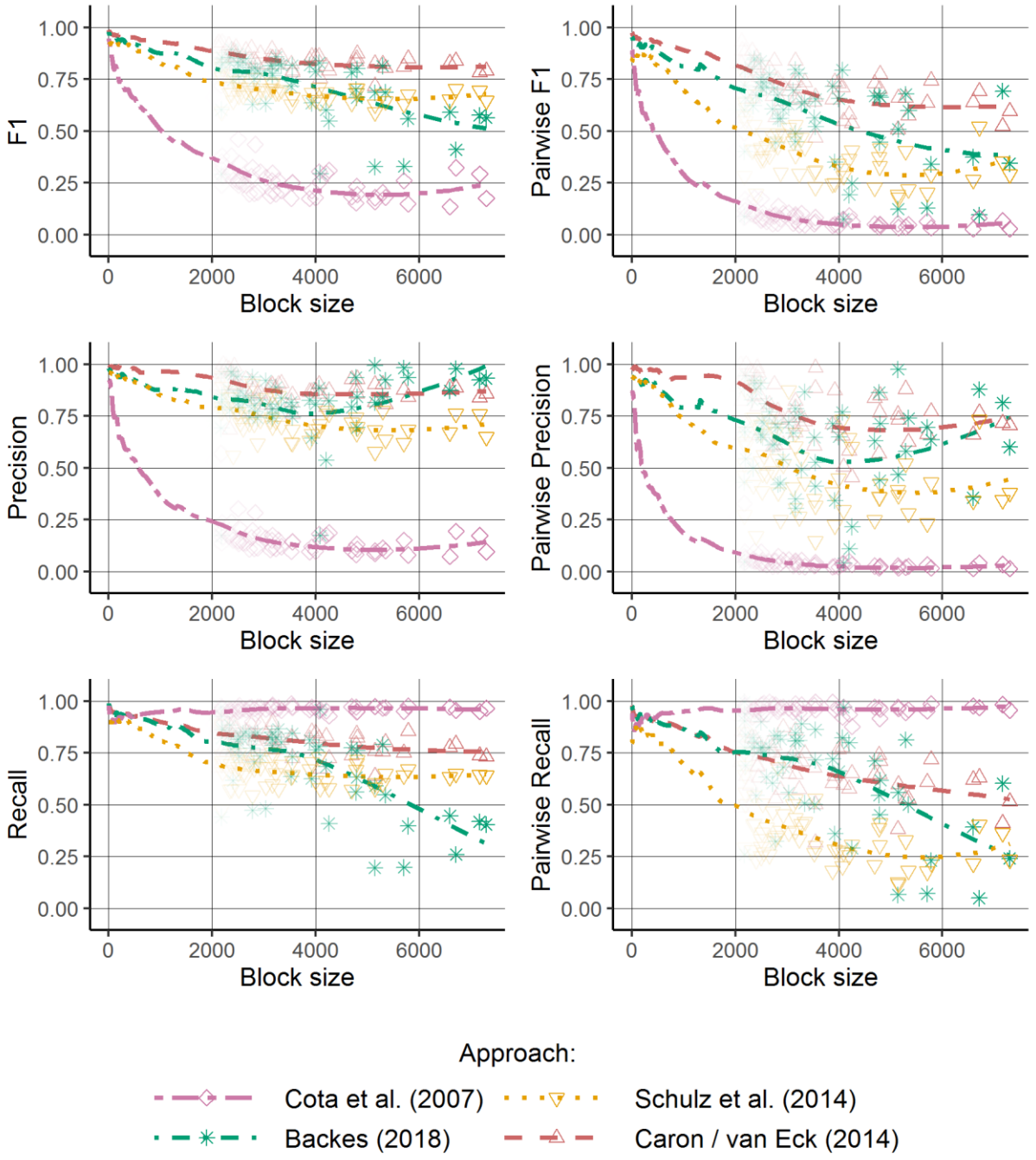


Figure 4-2. F1, precision and recall values for all approaches across block sizes using flexible thresholds (the best possible threshold/s/ is /are/ used for each block).
The lines show nonparametric regression estimates based on all blocks; the points show results for single blocks (only results for large blocks are displayed this way).

Similar to the approach of Cota et al. (2007), the difference between the original implementation and the one with flexible thresholds is rather small for the approach of Caron and van Eck (2014). However, the original implementation already uses different thresholds based on the block size classes. As the comparison with an implementation based on a constant threshold for all block sizes shows, this improves the results. Table 4-5 shows the evaluation results for the approach of Caron and van Eck (2014) with three different types of thresholds: a constant threshold for all blocks ("Constant"), the thresholds of the block size classes shown in Table

4-3 (“Block size classes”), and the optimal threshold for each single block (“Flexible”). These results show that the original implementation produces better results than those obtained using a constant threshold. This means that the somewhat rough partitioning between six block size classes allows for adequate differentiation with regard to the threshold and this strategy improves the disambiguation result compared to a constant threshold over all block sizes. In contrast, the strategy of specifying a threshold which is linearly dependent on the block size, as employed by the approach of Backes (2018a), is unable to find good thresholds over the complete range of block sizes. This is due mainly to a drop in the recall (together with an increasing precision) for large blocks. The thresholds chosen by the algorithm are thus too high for large blocks. Hence, a linear relationship between block size and threshold does not appear to be an adequate strategy for large blocks. The fitted thresholds for the approach of Caron and van Eck (2014) also confirm that a nonlinear relationship between block size and threshold may be more suitable. When using flexible thresholds instead of specifying them based on a linear relationship with the block size, the results for the approach of Backes (2018a) are close (even though with more variation among large blocks) to the results for the approach of Caron and van Eck (2014). This suggests that the approach of Backes (2018a) has the potential for producing good results if adequate thresholds are specified.

Table 4-5. Results for different types of thresholds for Caron and van Eck (2014)

Type of threshold	P_{pair}	R_{pair}	$F1_{pair}$	P_{best}	R_{best}	$F1_{best}$
Constant	0.690	0.741	0.714	0.879	0.880	0.880
Block size classes	0.831	0.787	0.808	0.916	0.885	0.900
Flexible	0.907	0.850	0.878	0.954	0.897	0.924

The results in Figure 4-2 and Table 4-5 demonstrate that the disambiguation quality can be improved if flexible thresholds dependent on the block size are specified. However, the specification of adequate thresholds is generally a nontrivial task, as it depends on the data at hand. Likewise, the thresholds proposed previously for the approaches examined in this paper do not correspond to the thresholds fitted with regard to our data set.

4.5.3 The influence of attributes considered for assessing similarities

Another important feature of disambiguation approaches is the set of the author mentions’ attributes they consider for assessing the similarity between author mentions. The different quality of the disambiguated data may result from considering different sets of attributes. For example, while Caron and van Eck (2014) included the attributes listed in Table 4-2, Schulz et al. (2014) only considered shared coauthors, shared cited references, shared citing papers, and self-citations. As less information was considered in the latter approach, this may be a reason why Caron and van Eck (2014) is better able to detect correct links between author mentions.

To get an idea of how important the set of attributes considered by the approaches is, we compared modified versions of the three approaches producing the best results in their original

versions. Using a subset of the originally proposed attributes for an approach is generally possible, simply by including these attributes as before and omitting the other attributes. However, it is not always similarly easy to include new attributes. The approach of Backes (2018a) is very flexible in this regard, because attributes (e.g., journal or subject) are weighted equally, and features (e.g., *Nature* or *Science* for the attribute "journal") are weighted automatically. Both types of weights could be easily applied to new attributes. In contrast, Schulz et al. (2014) and Caron and van Eck (2014) provide specific weights for each attribute. For these two approaches, it is not specified how new attributes can be weighted for calculating the similarity between author mentions, making them less flexible for the consideration of new attributes in the disambiguation process.

For our comparison, we disambiguated the data with the approach proposed by Caron and van Eck (2014) once more, but this time based on a reduced set of attributes, such that it corresponds to the attributes considered in the approach of Schulz et al. (2014). Furthermore, we disambiguated the data another two times with the approach proposed by Backes (2018a): in one case based on attributes similar to those considered by Schulz et al. (2014), in the other case based on attributes similar to those considered by Caron and van Eck (2014). In these two cases, the sets of attributes are not exactly the same, because self-citations cannot be included in the approach of Backes (2018a) in the same way as in the other two approaches. In the approach of Backes (2018a), similarities are calculated based on the features that two author mentions have in common for the same attributes.

For example, if the author names for cited references of two author mentions are represented by $R_1 = \{r_{11}, r_{12}, r_{13}, r_{14}\}$ and $R_2 = \{r_{21}, r_{22}, r_{23}\}$, respectively, the approach could consider the names occurring in both R_1 and R_2 for determining the similarity of the two author mentions. However, self-citations can only be detected by comparing the author names of cited references of one author mention with the name of the author itself of the second author mention. Such a comparison between two different attributes (here: author name and author names of cited references) is not intended in the original approach. There are no shared self-citations and the specificity of self-citations cannot be captured with the framework introduced by Backes (2018a) for calculating similarities between clusters of author mentions (we refrained from modifying this framework, which may be a possibility to include self-citations).

To keep the attribute sets comparable and still include self-citations in the approaches of Schulz et al. (2014) and Caron and van Eck (2014), we used information as close as possible in the approach of Backes (2018a) by including referenced author names instead of self-citations. We consider this choice to be appropriate. In the case that two of an author's mentions have self-citations to a third author mention of the same author, these mentions would also occur as shared referenced authors. Vice versa, if two author mentions share referenced authors, it is likely that self-citations are among these, because self-citations are usually overrepresented among cited references. An alternative to this choice of attribute sets would be to exclude self-citations and author names of cited references. However, our analyses show that these two alternatives (with

or without referenced authors and self-citations) produce similar results, and the conclusions are the same for both alternatives.

For each comparison and each approach, we separately specified the thresholds as described in section 4.3.2. The results of the outlined implementations are summarized in Table 4-6. The results show that differences between the approaches still exist. Characteristics of the approaches other than the set of attributes are therefore also relevant for the quality of an algorithm. In our analyses, the approach of Caron and van Eck (2014) produces the best results in any case, which indicates that the differentiation of block size classes for specifying thresholds and the weighting of attributes based on expert knowledge are appropriate concepts for disambiguating bibliometric data. Even though not as good as this approach, the approach of Backes (2018a) also produces good results in the comparisons. Its strategy to consider the specificity of particular features for determining the similarity of author mentions seems to be a promising approach, even if uniform weights are applied on the attribute level.

Table 4-6. Comparisons based on similar sets of attributes

Attribute set	Approach	$F1_{pair}$	$F1_{best}$
Schulz et al. (2014)	Schulz et al. (2014)	0.455	0.773
	Caron and van Eck (2014)	0.637	0.807
Schulz et al. (2014)	Schulz et al. (2014)	0.455	0.773
	Backes (2018a)	0.770	0.819
Caron and van Eck (2014)	Caron and van Eck (2014)	0.808	0.900
	Backes (2018a)	0.721	0.765

However, the results in Table 4-6 also reveal that the choice of attributes has a significant effect on the disambiguation quality. This can be concluded from the differences between the evaluation metrics for the approach of Caron and van Eck (2014) in its original implementation ($F1_{pair}$: 0.808, $F1_{best}$: 0.900), and its implementation used for the comparison with the approach of Schulz et al. (2014) ($F1_{pair}$: 0.637, $F1_{best}$: 0.807): The consideration of more attributes (the original implementation) produces better results. The importance of the choice of attributes also becomes obvious with regard to the results of the approach proposed by Backes (2018a). In this case, however, using more attributes does not necessarily produce better results: Using the same attributes as the approach of Schulz et al. (2014) produces better results than the original implementation (which is based on a larger set of attributes). The reason may be that some of the attributes considered in the original implementation have too much influence in the disambiguation procedure due to the uniform weights on the attribute level. Backes (2018a) also provides the possibility to apply different weights on the attribute level. This might be an alternative for improving the results when including the additional attributes. However, we did not consider this alternative, as the weights for the attributes are not specified automatically by the approach. They would have to be specified manually. Again, this suggests that not

only the choice of attributes, but also their weights, play a key role for the quality of disambiguation algorithms.

4.6 Discussion

In this study, we compared different author name disambiguation approaches based on a data set containing author identifiers in the form of ResearcherIDs. This allows a better comparison of different approaches than previous evaluations, because the comparisons in previous evaluations are generally based on different databases (which are scarcely comparable then). Our results show that all approaches included in the comparison perform better than a baseline that only uses a canonical name representation of the authors for disambiguation. The comparison in this study does not point to the recommendation of one approach for all situations that require a disambiguation of author names. It provides evidence of when which approach can produce good results – especially with regard to the size of corresponding name block sizes. Our analyses show that the parametrization of the approaches can have a significant effect on the results. This effect depends largely on the data at hand. Therefore, a proper implementation of an algorithm always has to take into account the characteristics of the data that has to be disambiguated. In the context of this study (based on its data set), the approach proposed by Caron and van Eck (2014) produced the best results.

Beyond the comparison of the original versions of the approaches, we also examined the role that the set of attributes – used by the different approaches – has on the results. As the approaches vary in the attributes they used for assessing the similarities between author mentions, differences in the results may rely on the choice of attributes. Our analyses show indeed that this choice has an effect on the results. Differences between the approaches, however, still remain when controlling for the set of attributes included. This means that other features of the approaches (e.g., how similarities are computed, or how similar author mentions are combined to clusters) also have an effect on the disambiguation quality. Based on these findings, we recommend that future research further examines the importance of single attributes and how they should ideally be weighted. The effect of the clustering strategy on the results might be also a topic for future research.

Regarding the evaluation of disambiguation approaches, we tested the results against author profiles from ResearcherID. As these profiles are curated by researchers themselves, the approaches are tested against human-based compilations of publications (i.e., compilations of those humans who are in the best position to reliably assign the publications to their personal sets). It would be interesting to compare the disambiguation approaches with other human-based compilations (e.g., ORCID) to see whether our results are still valid. We do not expect that the results will change significantly; we assume, however, that all human-based compilations are concerned with more or less erroneous records.

Understanding how author name disambiguation approaches behave is important to improve the applied algorithms and to assess the effect they have on analyses that are based on the

disambiguated data. A good understanding of this behavior is the basis for reliable bibliometric analyses at the individual level. It is clear that the same is true for any other unit (e.g., institutions or research groups) that is addressed in research evaluation studies.

References

Backes, T. (2018a). Effective unsupervised author disambiguation with relative frequencies. In J. Chen, M. A. Gonçalves, & J. M. Allen (Eds.), *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 203-212). ACM. https://doi.org/10.1145/3197026.3197036

Backes, T. (2018b). The impact of name-matching and blocking on author disambiguation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 803-812). ACM. https://doi.org/10.1145/3269206.3271699

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden* (pp. 79-86). Universiteit Leiden - CWTS.

Cota, R. G., Gonçalves, M. A., & Laender, A. H. F. (2007). *A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries* XXII Simpósio Brasileiro de Banco de Dados, João Pessoa.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic mMethods for author name disambiguation. *ACM SIGMOD Record*, *41*(2), 15-26.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2010). *Effective self-training author name disambiguation in scholarly digital libraries* Proceedings of the 10th annual joint conference on Digital libraries, Gold Coast, Queensland, Australia.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science and Technology*, *65*(6), 1257-1278. https://doi.org/10.1002/asi.22992

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review*, *32*. https://doi.org/10.1017/s0269888917000182

Hussain, I., & Asghar, S. (2018). DISC: Disambiguating homonyms using graph structural clustering. *Journal of Information Science*, *44*(6), 830-847. https://doi.org/10.1177/0165551518761011

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: a case of DBLP. *Scientometrics*, *116*(3), 1867-1886. https://doi.org/10.1007/s11192-018-2824-5

Kim, J. (2019). Scale-free collaboration networks: An author name disambiguation perspective. *Journal of the Association for Information Science and Technology*, *70*(7), 685-700. https://doi.org/10.1002/asi.24158

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology*, *67*(6), 1446-1461. https://doi.org/10.1002/asi.23489

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*(5), 1030-1047. https://doi.org/10.1002/asi.22621

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., & Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy*, *43*(6), 941-955. https://doi.org/10.1016/j.respol.2014.01.012

Liu, Y., Li, W., Huang, Z., & Fang, Q. (2015). A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, *66*(3), 634-644. https://doi.org/10.1002/asi.23183

Menestrina, D., Whang, S. E., & Garcia-Molina, H. (2010). Evaluating entity resolution results. *Proceedings of the VLDB Endowment*, *3*(1), 208-219.

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, *7*(4), 767-773. https://doi.org/10.1016/j.joi.2013.06.006

Newcombe, H. B. (1967). Record linking: The design of efficient systems for linking records into individual and family histories. *American Journal of Human Genetics*, *19*(3), 335–359.

On, B.-W., Lee, D., Kang, J., & Mitra, P. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. In M. Marlino (Ed.), *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 344–353). ACM. https://doi.org/10.1145/1065385.1065463

Schulz, C., Mazloumian, A., Petersen, A. M., Penner, O., & Helbing, D. (2014). Exploiting citation networks for large-scale author name disambiguation [journal article]. *EPJ Data Science*, *3*(11). https://doi.org/10.1140/epjds/s13688-014-0011-3

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1-43. https://doi.org/10.1002/aris.2009.1440430113

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3), 1-29. https://doi.org/10.1145/1552303.1552304

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, *101*(3), 1955-1972. https://doi.org/10.1007/s11192-014-1283-x

Wu, J., & Ding, X.-H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics*, *96*(3), 683-697. https://doi.org/10.1007/s11192-013-0978-8

Zhu, J., Wu, X., Lin, X., Huang, C., Fung, G. P. C., & Tang, Y. (2017). A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering. Scientometrics, 114(3), 781-794. https://doi.org/10.1007/s11192-017-2611-8

5 Gender and scientific output: How productivity, citation impact and journal prestige differ between female and male researchers

Alexander Tekles

Abstract

Gender differences in scientific output have been discussed as one possible reason for the underrepresentation of women in science. Therefore, it is important to understand the extent and the contexts in which these gender differences exist. This study aims to contribute to this understanding by empirically analysing gender differences in scientific output in terms of productivity, citation impact and journal prestige. The analyses are based on Scopus data and focus on the cohort of researchers who started publishing in 1999. Their publications were tracked for up to 20 years. The results demonstrate that male researchers publish more papers than female researchers and that this difference increases over the course of scientific careers. However, after controlling for the disciplines in which researchers publish, the productivity difference declines and even disappears among researchers with short careers. By contrast, female researchers achieve higher citation impact and publish in more prestigious journals than male researchers over the course of their careers, especially among researchers with short careers. The results suggest that many women with high potential leave the science system early in their careers and that the choice of bibliometric indicators may differentially affect the evaluation of female and male researchers.

5.1 Introduction

Many studies have provided evidence for various forms of gender differences in science. For example, empirical analyses suggest that female and male researchers differ with regard to productivity (Halevi, 2019), citations received (Larivière et al., 2013), journal prestige (Larivière & Sugimoto, 2017), collaborations (Zeng et al., 2016), mobility (de Kleijn et al., 2020), funding (Witteman et al., 2019) and chances to be hired (Moss-Racusin Corinne et al., 2012). Such gender differences have been discussed as possible reasons for the underrepresentation of women in science. While the share of women in science has generally increased over the past few decades, there is consistent empirical evidence that women are still underrepresented in science (de Kleijn et al., 2020). Therefore, it is important to understand the extent and the contexts in which gender differences exist and may contribute to the underrepresentation of women in science. This study aims to contribute to this understanding by empirically analysing gender differences in scientific output in terms of productivity (i.e., the number of papers that a researcher has published), citation impact (i.e., the number of citations that a researcher's publications have received) and journal prestige (i.e., the impact of the journals in which a researcher has published).

Scientific output is not only a marker of scientific success in itself, but it is also relevant for researchers' careers. One reason is that a researcher's scientific output can affect the chances of being hired (Berenbaum, 2019; Jungbauer-Gans & Gross, 2013; Thelwall, 2020). Scientific output may also affect the likelihood of pursuing a scientific career in the first place, as past success in the form of scientific output may increase a researcher's motivation to remain in academia and the perceived chances of succeeding in a scientific career. Therefore, it is important to have an accurate picture of gender differences in scientific output.

Existing studies on gender differences in scientific output have used different methodological approaches to measure them, but they often fail to justify why a particular approach was selected, making it difficult to compare and interpret many of these studies. The current study discusses what the different methodological approaches actually measure and how suitable they are for measuring gender differences in scientific output. The most suitable approaches were applied in the empirical analyses of this study. Furthermore, mediating factors between gender and scientific output were controlled for that other studies often did not control adequately. In particular, female and male researchers with a similar academic age (i.e., the time elapsed since they began their academic careers as actively publishing researchers) and who have published in similar disciplines were compared. Controlling for these factors is necessary because scientific output is only comparable within disciplines and between researchers who have been active in science for a similar amount of time.

The empirical analyses in this study are based on Scopus data. These include the Scopus Author ID, a reliable identifier for researchers, which was used to generate bibliometric data at the individual level. The scientific output of all researchers who first published in 1999 was tracked

over the course of up to 20 years. The analyses focus on this cohort because the researchers' publication history could be reliably identified for it, while also observing their publications over a relatively long time span (see Section 5.3.1). To control for the researchers' academic age, the scientific output of female and male researchers was compared over their careers, which also allows to examine possible variations in gender differences over the researchers' academic age. Since the early phase of a scientific career is crucial for its further development, such variations may be an additional reason for the underrepresentation of women in science, beyond mere level differences that are constant over the researchers' academic age. The researchers' disciplines were controlled for by measuring pairwise similarities between the researchers based on the disciplines in which they have published. This methodological approach provides more flexibility with the identification of a researcher's disciplines and an alternative to ordinary field classification systems that assign researchers (or other entities such as journals or papers) to one or a few particular disciplines. Furthermore, the researchers' total career length was considered in the empirical analyses. While academic age indicates the time until the point at which a researcher's scientific output is measured, the total career length indicates how long a researcher has been active in science (even beyond the point in time at which the output is measured). In other words, the scientific output can be observed at different academic ages for a researcher, but the career length is fixed for each researcher. Considering the total career length allows to control for a potentially higher selectivity of female researchers who successfully pursued a scientific career.

5.2 Scientific output and gender

Existing studies on gender differences in scientific output have used different approaches to measure scientific output. Sections 5.2.1 to 5.2.3 discuss these approaches and summarise the empirical results reported in the studies. Besides these empirical results, the literature provides some theoretical arguments for the existence of gender differences in scientific output, such as a possible discrimination against female researchers in the assessment of scientific work (Helmer et al., 2017), more teaching responsibilities for female than male researchers (Thelwall, 2018), gender differences in the aims researchers (Zhang et al., 2021), gender-specific collaboration patterns (Jadidi et al., 2018) or differences in access to resources (Duch et al., 2012). The empirical analyses in this study are rather descriptive and cannot test these mechanisms. However, some important mediating mechanisms were controlled for, as they may produce gender differences merely due to composition effects. These mechanisms are discussed in Section 5.2.4. In particular, only researchers who have published in similar disciplines, with a similar total career length and at a similar academic age were compared. Gender-specific selection processes may result in different gender distributions across disciplines, academic age and career length, while these factors may also affect scientific output.

5.2.1 Productivity

Many studies have already analysed gender differences in scientific productivity in terms of number of published papers. A common methodological approach in these studies is to compare the number of female-authored papers and the number of male-authored papers over a specific time period (and in a particular discipline). Studies that applied this approach have consistently found that most papers are male-authored (see Halevi, 2019, for an overview). However, simply counting the number of female- or male-authored papers only measures the extent to which female and male researchers contribute to science over a certain time period (and in a particular discipline). Since the primary focus of this study is to assess the productivity of female and male researchers rather than their overall contributions to the science system, the empirical analyses must also examine productivity at the individual level. Consequently, the number of papers per researcher rather than the overall number of female- and male-authored papers must be considered.

Until recently, such analyses at the individual level have been difficult to conduct due to a lack of individual-level bibliometric data. As a consequence, most studies that examine productivity in terms of papers published per researcher are restricted to a specific and usually small set of researchers for which reliable individual-level data are available (e.g., Aaltojärvi et al., 2008; Ebadi & Schiffauerova, 2016; Fox, 2005; Mauleón & Bordons, 2006; Mayer & Rathmann, 2018; Paik et al., 2014; Raj et al., 2016; van Arensbergen et al., 2012). However, the emergence of algorithms for disambiguating author names in bibliometric data and a growing number of datasets that already include disambiguated data have made it easier to conduct empirical analyses at the individual level (Tekles & Bornmann, 2020). For example, Boekhout et al. (2021) and de Kleijn et al. (2020) used the author identifier provided by Scopus, while Huang et al. (2020) used different bibliographic databases with author identifiers to analyse gender differences in productivity.

Most existing evidence on gender differences in productivity at the individual level also suggests a consistent productivity gap, with male researchers publishing more papers per year than female researchers. For example, Ceci et al. (2014) summarised several studies that found such gender differences in individual productivity in STEM fields. Moreover, recent studies by Boekhout et al. (2021) and de Kleijn et al. (2020) found gender differences in individual productivity based on a comprehensive dataset covering various disciplines. In contrast to the vast majority of studies on gender differences in productivity, Huang et al. (2020) found no such differences in productivity after controlling for career length (by measuring a researcher's productivity using the average number of papers published per year). Other studies on gender differences in individual productivity are mostly restricted to specific fields or relatively small datasets, but they have consistently found that male researchers publish more papers than female researchers (e.g., Aaltojärvi et al., 2008; Akbaritabar & Squazzoni, 2020; Fox, 2005; Mauleón & Bordons, 2006; Paik et al., 2014; Raj et al., 2016).

5.2.2 Citation impact

In contrast to the empirical results regarding gender differences in productivity, findings on gender differences in the citation impact of female and male researchers' publications are mixed. To assess gender differences in citation impact, many studies have compared the average number of citations between female- and male-authored papers. While some studies that followed this approach have found that male-authored papers achieve a higher citation impact than female-authored papers (Andersen et al., 2019; Chatterjee & Werner, 2021; Zhang et al., 2021), others have found that female-authored papers achieve a similar or higher citation impact than male-authored papers (Ceci et al., 2014; Halevi, 2019; Lynn et al., 2019; Thelwall, 2018). Comparing citation impact at the paper level allows to assess how publications are cited based on the authors' gender, but it does not allow to measure gender differences at the individual level (i.e., how many citations a researcher receives).

However, it is possible to assess gender differences in citation impact at the individual level by comparing the total number of citations for female and male researchers (i.e., the sum of citations received for all of a researcher's papers). Fewer studies have used this approach because it requires disambiguated data at the individual level (in contrast to simply comparing female- and male-authored papers, which does not require disambiguated data). The scarce evidence that does exist suggests that male researchers have a higher total number of citations (Huang et al., 2020), which is unsurprising given the gender differences in productivity. The total number of citations received by a researcher's papers can be interpreted as an indicator that measures two things at once: the number of papers that a researcher has published over the entire career and the average impact of the researcher's papers. To measure citation impact only, the average number of citations per paper rather than the total number of citations must be compared between female and male researchers. Empirical results based on this approach do not suggest gender differences in citation impact (de Kleijn et al., 2020; Huang et al., 2020).

5.2.3 Journal prestige

Although the use of journal metrics in the evaluation of researchers has been repeatedly criticised (Berenbaum, 2019), the impact of the journals (i.e., journal prestige) in which a researcher publishes remains relevant in evaluation contexts (McKiernan et al., 2019). Compared to research on gender differences in productivity and the citation impact of publications, less empirical evidence is available on whether female and male researchers differ with regard to the impact of the journals in which they publish. Some evidence suggests that female researchers tend to publish in lower-impact journals than male researchers (Joanis & Patil, 2022; Larivière & Sugimoto, 2017; Mayer & Rathmann, 2018), especially in prestigious authorship positions (Bendels et al., 2018).

Other studies have focused on possible mechanisms that can lead to gender differences in journal prestige. For example, several studies have compared acceptance rates for journal submissions to test whether the peer review process may be influenced by a gender bias. Such a gender

bias could contribute to gender differences in journal prestige if it especially influences the review process in top journals. However, empirical evidence on a possible gender bias in the journal peer review process is mixed (Kern-Goldberger et al., 2022; Squazzoni et al., 2021). Gender differences in journal prestige could also be attributed to differences in submission rates to high-impact journals between female and male researchers, which some studies have reported empirical evidence for (e.g., Breuning & Sanders, 2007; Teele & Thelen, 2017).

5.2.4 The role of discipline, academic age and total career length

When analysing gender differences in scientific output, the disciplines in which researchers have published should be controlled for. Different disciplines exhibit different publication rates (Bornmann, 2019), average citation counts (Waltman & van Eck, 2019) and, by extension, different levels of average journal impact. At the same time, there is a gender-specific segregation of researchers into disciplines (Holman et al., 2018; Tekles et al., 2022; Thelwall et al., 2020; West et al., 2013). For example, female researchers may be predominantly active in discipline A, in which researchers usually only publish a few papers over a given time period, while male researchers may be predominantly active in discipline B, in which researchers usually publish more papers. Consequently, simply comparing female and male researchers across different disciplines could result in gender differences with regard to the number of published papers only due to differences in the disciplines' publication cultures.

Among the existing studies on gender differences in scientific output, there are two different strategies to control for discipline at the individual level. One strategy is to focus on a particular discipline, which automatically implies to control for discipline (e.g., Aaltojärvi et al., 2008; Akbaritabar & Squazzoni, 2020). Another strategy is to assign researchers to a discipline and use this information in the analyses (Boekhout et al., 2021; Huang et al., 2020; Mayer & Rathmann, 2018). In the latter case, researchers who have published in multiple disciplines must be excluded from the analyses or assigned to the discipline that they have predominantly published in, which could distort the data. When measuring citation and journal impact, a field classification system can be used to control for discipline (e. g. the ASJC provided by Scopus, which assigns each journal to one or a few disciplines). Although this approach reduces the problem of differences in scientific output between disciplines, field classification systems may be too broad to identify all of these differences. Furthermore, field-normalised indicators may not adequately account for multidisciplinary research output. To mitigate these issues, this study uses an alternative approach to control for discipline at the individual level (see Section 5.3.2).

Besides the disciplines in which researchers publish, their academic age should also be controlled for. Due to the higher probability of female researchers for leaving science throughout their scientific careers compared to male researchers (Huang et al., 2020; Jadidi et al., 2018), it can be assumed that male researchers have a higher average academic age than female researchers. At the same time, a researcher's output may increase over the course of the researcher's

career due to resources, competencies and other advantages accumulated over time (Boekhout et al., 2021; van den Besselaar & Sandström, 2017). In this case, male researchers would have a higher level of scientific output than female researchers simply because they have a higher academic age. For example, if male researchers have, on average, been in science for longer than female researchers, they have been able to gain more experience, which may increase their productivity even in the absence of other effects of gender on productivity. In this study, the empirical analyses account for this factor by separately assessing gender differences for each year of a researcher's career (academic age).

This approach not only allows for academic age to be controlled for, but it also enables to analyse how gender differences change over the course of scientific careers. While there is considerable evidence on gender differences in the level of scientific output (especially for productivity and citation impact), few studies have examined how these gender differences change with academic age. Variations over academic age may occur even if there are no overall gender differences (as some studies have suggested for citation impact). If there are overall gender differences (as studies have suggested for productivity and journal prestige), there may be even larger gender differences during certain career phases and smaller gender differences during other career phases.

For example, van den Besselaar and Sandström (2016) calculated differences in productivity and citation impact between female and male researchers among 400 social science researchers in the Netherlands at the beginning of their careers and 10 years later. Their results indicate that gender differences in productivity increase over the course of scientific careers; specifically, the productivity of male researchers develops stronger than that of female researchers. However, the authors found no gender differences in citation impact at either point in time. The methodological approach of van den Besselaar and Sandström (2016) provides an empirical analysis of linear trends in gender differences over scientific careers, but it does not allow to draw any conclusions about non-linear trends. By contrast, Boekhout et al. (2021) recently reported results on the development of gender differences in productivity over time that would also allow to observe non-linear trends in gender differences. However, their results only indicate a linear trend of increasing differences in productivity over the first 16 years of scientific careers, which aligns with the findings of van den Besselaar and Sandström (2016).

Besides the scarce empirical evidence, there are some theoretical arguments supporting the hypothesis that gender differences in scientific output vary with academic age. For example, young researchers may be more likely to face gender bias in evaluations than eminent researchers, who are judged on their accomplishments rather than their gender. In the early phases of their career, female researchers may also be affected by career absences and family responsibilities more than male researchers (Jungbauer-Gans & Gross, 2013; van den Besselaar & Sandström, 2016). These exemplary mechanisms illustrate that gender differences in scientific output may vary with academic age. Separately measuring gender differences for each academic age allows to empirically examine this relationship.

To reliably assess how gender differences change with academic age, a researcher's total career length must also be controlled for. If female researchers face higher costs of pursuing a scientific career, it can be expected that female researchers with long careers are very selective with regard to their chances to be successful (e.g., because they are especially motivated and skilled). By contrast, male researchers with long careers may be less selective. Differences in selectivity between female and male researchers would be smaller for short careers because the costs of pursuing a scientific career play a less important role for short careers than for long careers. In this case, variations in gender differences with academic age and career length would interfere. At an advanced academic age, female researchers would have a higher output (i.e., productivity, citation impact and journal prestige) than female researchers at an early academic age because only the selective group of female researchers with long careers would reach an advanced academic age. While this may also be true for male researchers, the pattern may be less pronounced for them, assuming that they are less selective than female researchers with increasing career length. Thus, variations in gender differences in scientific output with academic age would (partly) be attributable to differences in the selectivity of researchers with long careers.

5.3 Data and methods

5.3.1 Data

All empirical analyses in this study are based on data from the Scopus in-house database of the Competence Centre for Bibliometrics (https://bibliometrie.info/). This dataset contains bibliometric data on papers published between 1996 and 2021. Scopus is not only one of the most important data sources in bibliometrics (Visser et al., 2021) but also provides an identifier for researchers (the Scopus Author ID), which was used to generate data at the individual level. The Scopus Author ID relies on an author name disambiguation algorithm developed by Scopus. This algorithm aims to resolve synonyms (one researcher publishing under different names) and homonyms (different researchers with identical names) among author names to accurately determine the publication sets of researchers. The implementation details of the disambiguation algorithm are not disclosed, but evaluation results published by Scopus (Baas et al., 2020) and other analyses (Aman, 2018; Kawashima & Tomizawa, 2015; Reijnhoudt et al., 2014) suggest that the algorithm produces reliable author identifiers.

In the current study, only researchers who have published their first paper in 1999 were considered in the empirical analyses. Including earlier publication years would have resulted in a high proportion of author profiles for which the starting year of their publishing career could not be reliably identified. Researchers may have already published in 1995 or earlier (which is not covered by the dataset), but not in the first year(s) after 1995. By restricting the data to researchers whose first publication in the dataset is from 1999, these researchers would need to have a gap of at least three years to falsely assume that 1999 is also the year of their first publication. Thus, the approach can be assumed to exclude most falsely identified starting years of researchers' careers. The results of Boekhout et al. (2021) support this assumption, who also used

Scopus data covering publications from 1996 onwards and the Scopus Author ID to identify researchers. In their analyses of the share of female and male researchers among researchers with their first publication, they found that the time trend of these shares remains relatively stable from 1999 onwards but considerably differs in the years before.

The last publication year considered for the empirical analyses in Section 5.4 is 2018, allowing for a citation window of at least three years, which is necessary to reliably measure a paper's citation impact (Bornmann, 2019; Wang, 2013). Consequently, up to 20 years of a researcher's career are covered in the data. Besides these restrictions regarding publication years, author profiles with less than three publications were excluded because many of them can be assumed to be the result of errors by Scopus' author name disambiguation algorithm. This approach is in line with the analyses of Boekhout et al. (2021). When interpreting the results, it should be born in mind that this approach excludes researchers who have never published or only published a few papers, which is especially the case among researchers with short careers.

To determine the authors' gender, an open-source application was used to assign a gender to first names (Studer, 2012). This application also allows the country of origin to be considered for determining a person's gender. Since the country of origin is not available in the Scopus data, the authors' affiliation was used as a proxy for this purpose. Using the affiliations may have led to wrong assumptions about the country of origin in some cases, but it probably increased the reliability of the gender assignments in most cases. If a first name is associated with a different gender in other countries, the application classifies this name as probably female or male for the corresponding country. These cases were also included in the analyses. The application only distinguishes between female and male names, so that the empirical analyses are restricted to a binary concept of gender. For each researcher, the gender was determined for all name-country pairs that could be derived from the data (a researcher may have different names or affiliations due to homonyms or changes in affiliation). No gender was assigned in cases with unisex names or if only the initials of the first name were given. A gender was only assigned to a researcher (i.e., the researcher was included in the analyses) if the same gender has been assigned to all of the name-country pairs for which a gender could be assigned. Overall, the gender could be inferred for 72,992 researchers (25% of all researchers in the data); of these, 24,592 (34%) were classified as female and 48,400 (66%) as male.

5.3.2 Measuring gender differences in scientific output

The indicators for the three dimensions of scientific output described in Section 5.2 were calculated based on the bibliometric information in the Scopus data. A researcher's productivity was measured in terms of the number of papers that the researcher has published (up to a certain point in time). The citation impact of a researcher's papers was measured based on the papers'

citation percentiles, according to the approach of Hazen (1914). These percentiles are calculated using the formula

$$P_H = 100 \times \frac{i - 0.5}{n}$$

where i is the paper's rank within the set of papers published in the same field and year, and n is the number of papers in this set (Bornmann & Williams, 2020). For papers with identical citation counts, an average rank is assigned to all of them so that they have the same value for i. A P_H of 70 means that (approximately) 70% of the papers in the same discipline and publication year received fewer citations than the focal paper. The normalisation with regard to discipline and publication year allows to compare the citation impact of papers from different disciplines and years. Otherwise, such a comparison would not be meaningful because different citation counts can be expected for different disciplines and publication years (Hicks et al., 2015).

To measure the impact of the journals in which a researcher has published, the CiteScore was used. The CiteScore is an indicator developed and used by Scopus to measure the average citation impact of the papers published in a particular journal and year. For a given journal and publication year t, the CiteScore is defined as the average number of citations received by all articles, reviews, conference papers, book chapters and data papers published between $t - 3$ and t over the same time period. A paper's CiteScore was determined by the journal in which it has been published. The average CiteScore of a researcher's papers was then used as an indicator for the prestige of the journals in which the researcher has published. All indicators used in this study were calculated based on a full counting approach, which means that each of a researcher's publications was equally weighted, regardless of the number of co-authors.

To assess gender differences in scientific output over the course of scientific careers, the aforementioned indicators were calculated separately for each year of a researcher's career, starting with the first year in which a researcher has published. For most of the analyses, the indicators were calculated based on all publications until a certain academic age. This approach was used because it can be assumed to be more relevant for the career progression of researchers than the alternative approach to calculate the indicators based only on the publications in one particular year. For example, all of a researchers' publications are usually considered in hiring decisions rather than only works published in the previous year.

The researchers' total career length was controlled for by separately analysing researchers with different career lengths. The last publication year among a researcher's papers in the data was assumed to also be the last year of the career if the researcher has not published in a journal indexed by Scopus for at least three years after this date. Thus, the total career length was only inferred for researchers who have published their last paper in 2018 or earlier (i.e., within 20 years after their first publication). For researchers with longer careers, it can only be concluded that they have been actively publishing for more than 20 years and no further differentiation is possible with regard to career length.

The disciplines in which researchers have published were controlled for by matching female and male researchers who have published in similar disciplines. The ASJC journal classification provided by Scopus was used for this purpose, which assigns each journal in the database to one or more out of 334 disciplines. The similarity between two researchers was calculated based on the share of papers that a researcher has published in each discipline over the entire career. These shares were calculated based on a full counting approach: if 10% of a researcher's papers are assigned to a particular discipline, the share for this discipline is 10%, even if the papers are assigned to multiple disciplines. This approach results in one vector for each researcher, in which each element indicates the share of papers that the researcher has published in a discipline. The cosine similarity between the vectors of two researchers was used as the similarity between the researchers with regard to their disciplines. Based on this measure of similarity, a kernel matching approach was applied, with the researchers' gender as the treatment variable and the output indicators as outcome (Bittmann et al., 2021).

This was achieved by calculating a counterfactual outcome (output indicator) for each researcher based on the outcomes of researchers of the opposite gender and with the same total career length. The matched researchers were weighted differently in the calculation of the counterfactual outcome for a given researcher: the more similar a researcher of the opposite gender, the more weight this matched researcher had in the calculation of the counterfactual outcome. The counterfactual outcome can be interpreted as the output indicator value expected for a researcher of the opposite gender who has published in similar disciplines. The weights were calculated by applying the Epanechnikov kernel to the similarities between the researchers. Based on the factual and counterfactual outcomes for each researcher, the average treatment effects (i.e., gender differences when controlling for discipline) were calculated for each output indicator and career length.

5.4 Results

In a first step, gender differences in scientific output over the researchers' entire careers were analysed. Figure 5-1 shows the distribution of the three output indicators described in Section 5.3.2 for female and male researchers in terms of quartiles. The distributions are shown separately for each total career length. For example, Figure 5-1A shows that the quartiles for the number of papers that female and male researchers with a career length of 14 years have published over their careers are nearly identical with a median of eight papers. The blue lines in Figure 5-1 show the differences between the median number of papers published by female and male researchers with a particular career length. Overall, male researchers have published more papers than female researchers. This difference is rather small for career lengths of up to 20 years with a maximum difference of two in the median number of published papers for researchers with career lengths between 15 and 20 years. Hence, male researchers with a career length between 15 and 20 years have, on average, published two papers more than female researchers with the same career length. For researchers with longer careers (more than 20 years),

a considerably larger gender difference can be observed: over the first 20 years of their careers, male researchers have published nearly 10 papers more than female researchers in this group, on average.

To interpret this result, it is important to note that the difference in productivity was calculated based on the total number of papers over the first 20 years of a researcher's career. Even if the difference in the number of papers published per year (rather than the number of papers over a researcher's entire career) is constant over career lengths, a larger difference in the total number of papers can be expected for long careers than for short careers. For example, if male researchers publish an average of 1.2 papers per year and female researchers publish an average of one paper per year, the absolute difference amounts to one paper after five years but two papers after 10 years. Thus, the gender difference in the total number of published papers automatically increases over time even if the gender difference in average productivity per year remains constant.

In contrast to productivity, both citation impact and journal prestige are higher for female researchers than for male researchers. This result is relatively constant across different career lengths, which suggests that female researchers with long careers are not more selective than male researchers with long careers, or vice versa.

So far, only the researchers' total output over their entire careers was considered, which does not allow for an analysis of possible changes in gender differences with academic age. Therefore, gender differences in scientific output were separately analysed for each academic age in the next step of the analyses. Figure 5-2 shows the differences in the arithmetic mean of the three output indicators between female and male researchers for each academic age. These differences are separately plotted for short careers (up to 20 years; green dashed line) and long careers (more than 20 years; blue solid line). It should be noted that, for short careers, researchers could only be considered up to their career length, which implies that the set of researchers changes with academic age. For example, for the academic age of five years, all researchers who have actively published for at least five years could be considered. These researchers are a superset of all researchers who have actively published for at least 10 years and could be considered for the academic age of 10.

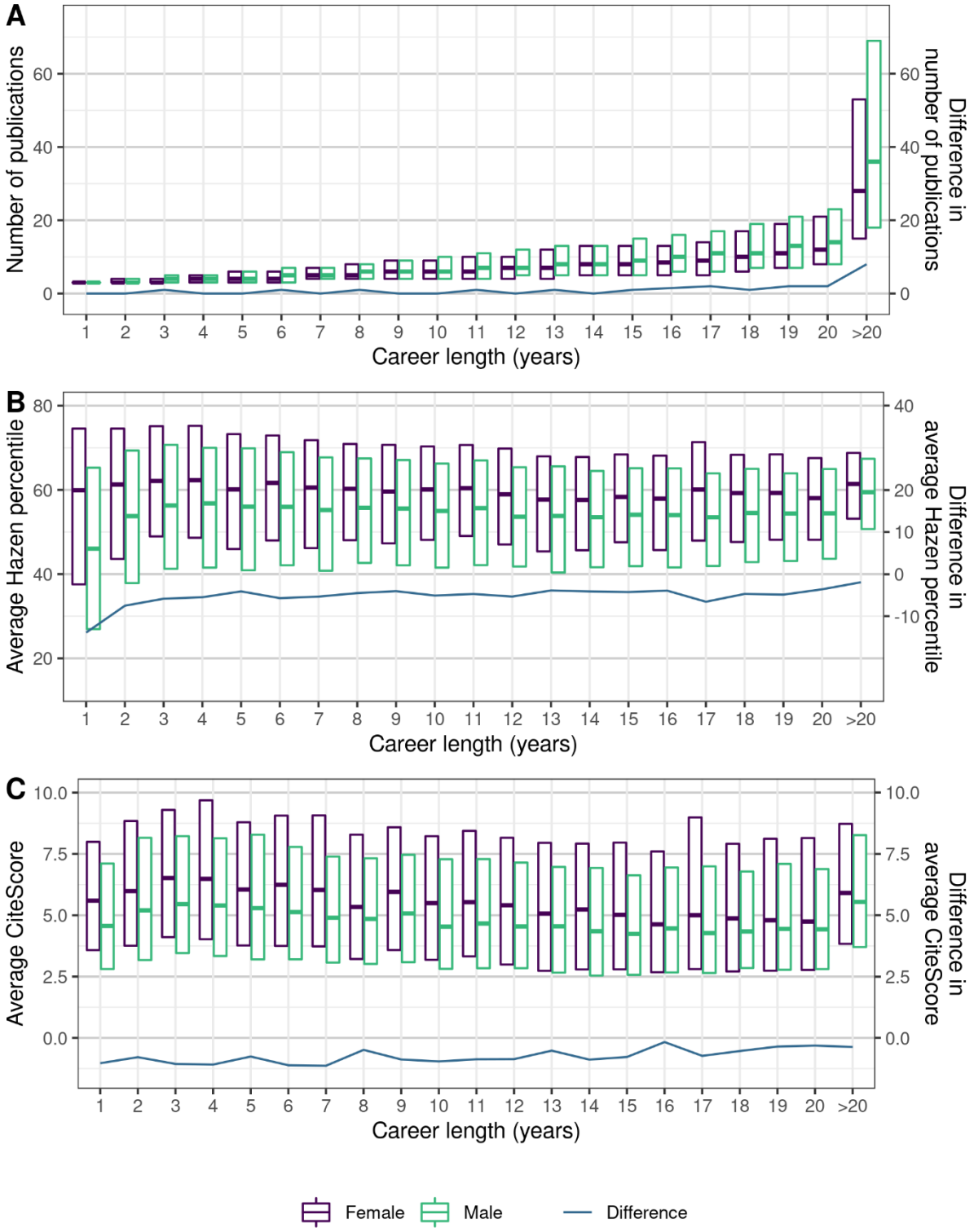


Figure 5-1. Scientific output over entire careers, separately for different career lengths. For career lengths larger than 20, the output was measured only over the first 20 years of a researcher's career. Scientific output was measured in terms of the total number (A), the average Hazen percentile (B), and the average CiteScore (C) of papers published over a researcher's career. The boxes show the lower quartile, median, and upper quartile for female and male researchers. The lines show the differences between the medians of male and female researchers (positive values indicate higher output for male researchers).

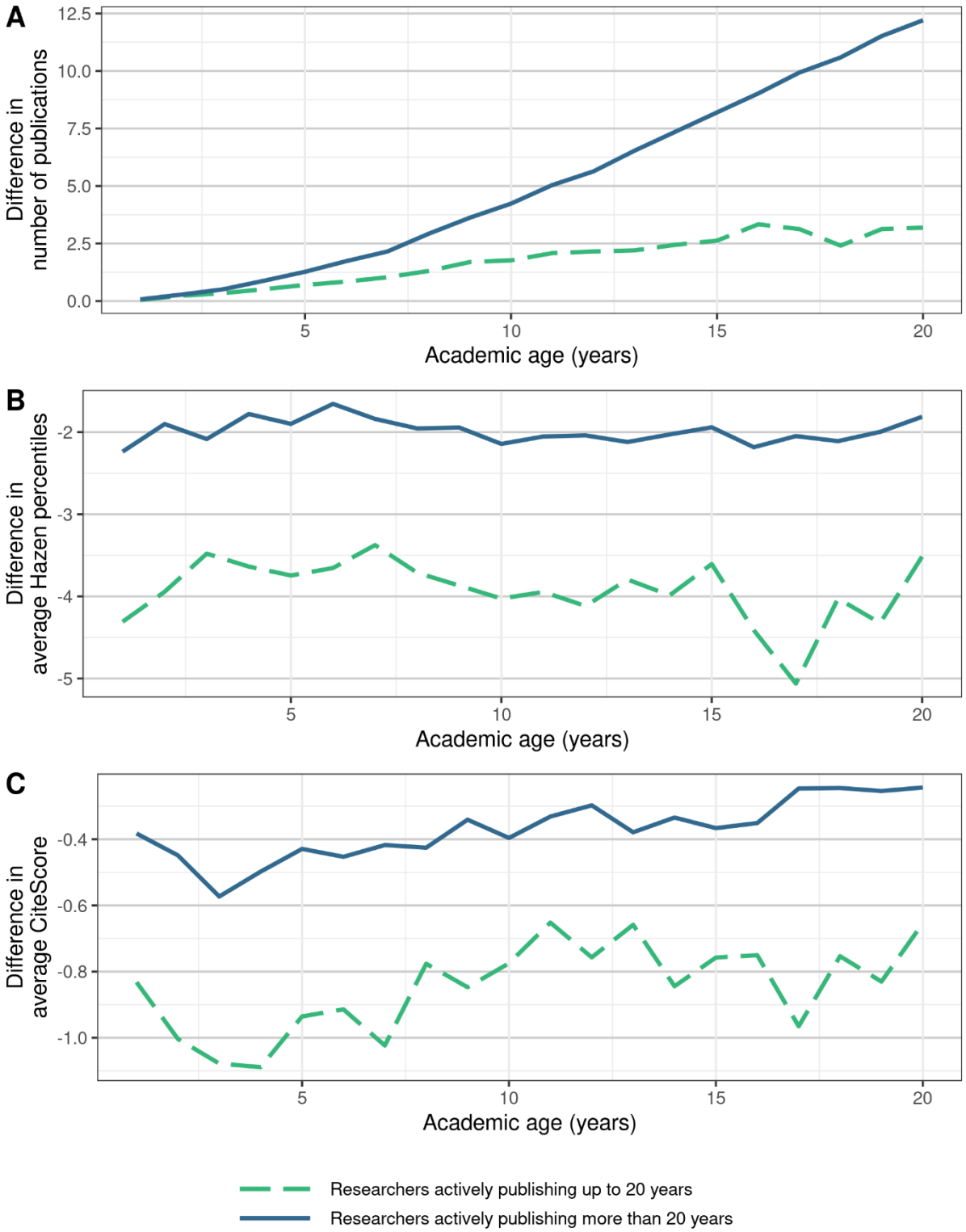


Figure 5-2. Differences in scientific output over academic age.
Scientific output was measured in terms of the total number (A), the average Hazen percentile (B), and the average CiteScore (C) of papers published until a particular academic age. Differences were calculated between the arithmetic means of male and female researchers (positive values indicate higher output for male researchers). The lines show the differences for the researchers who have published their last paper within 20 years after their first publication (green, dashed) and for researchers who have been actively publishing for more than 20 years (blue, solid).

Only differentiating between short and long careers is a rather general approach, and it could be argued that career lengths should be disaggregated even further. However, only differentiating between short and long careers allows the results to be presented more concisely than with a more granular approach. Separately analysing researchers for each career length up to 20 years shows that there are no clear patterns in the gender differences with regard to career length (up to 20 years): the differences are not consistently smaller or larger for increasing career length (see Figure 5-5). Thus, it can be assumed that pooling all researchers with career lengths of up to 20 years does not distort the effect of academic age on gender differences in scientific output. For researchers with long careers, a further differentiation is not possible due to the data restrictions mentioned in Section 5.3.1. However, researchers who have actively published for more than 20 years usually have an established scientific career and can thus be regarded as successful in pursuing a scientific career. Therefore, only differentiating between short and long careers appears to be an appropriate approach.

With regard to productivity, male researchers once again have published more papers than female researchers in all subgroups (i.e., career lengths and academic ages). Furthermore, this difference tends to widen over the course of the researchers' careers. For short careers, the difference is smaller than one paper within the first five years of a researcher's career. This difference plateaus after the academic age of 15 years and male researchers have, on average, published a maximum of approximately three papers more than female researchers. For long careers, the difference in productivity between female and male researchers is generally larger and increases more rapidly than for short careers. Hence, male researchers have published more papers than female researchers, especially among researchers with long careers. In summary, the results confirm a generally higher productivity among male researchers and this difference tends to increase over the course of the researchers' careers.

As for the interpretation of Figure 5-1A, it should be noted that the difference in the cumulative number of papers until a particular academic age is plotted, which should not be confused with the number of papers only published at a particular academic age. Therefore, the observed gender difference in productivity increases not only if the publication rate of male researchers increases, but also if they consistently publish more papers than female researchers, regardless of changes in the publication rate. The results shown in Figure 5-3B confirm this interpretation. For short careers, the difference in the number of papers published by a researcher at a particular academic age does not linearly increase, but follows a slight inverse u-shape around 0.1 papers per year that male researchers publish more than female researchers. For long careers, this difference increases until the academic age of 13 years and then reaches a plateau, whereas the difference in the cumulative number of papers increases even after this academic age.

The results for citation impact and journal prestige over the researchers' total career length (see Figure 5-1) are generally confirmed when analysing gender differences over academic age. Female researchers have a higher average citation impact and published in more prestigious journals than male researchers. This pattern remains relatively stable over academic age. For long

careers, the gender differences in citation impact and journal prestige are smaller than for short careers. Thus, female researchers have a higher citation impact and publish in more prestigious journals than male researchers especially among researchers with short careers (i.e., if they leave the science system as actively publishing researchers within 20 years after their first publication). Assuming that researchers with a high citation impact or researchers who publish in prestigious journals early in their career are especially valuable for the science system, this result suggests that the science system loses disproportionately many women with high potential in their early career phase.

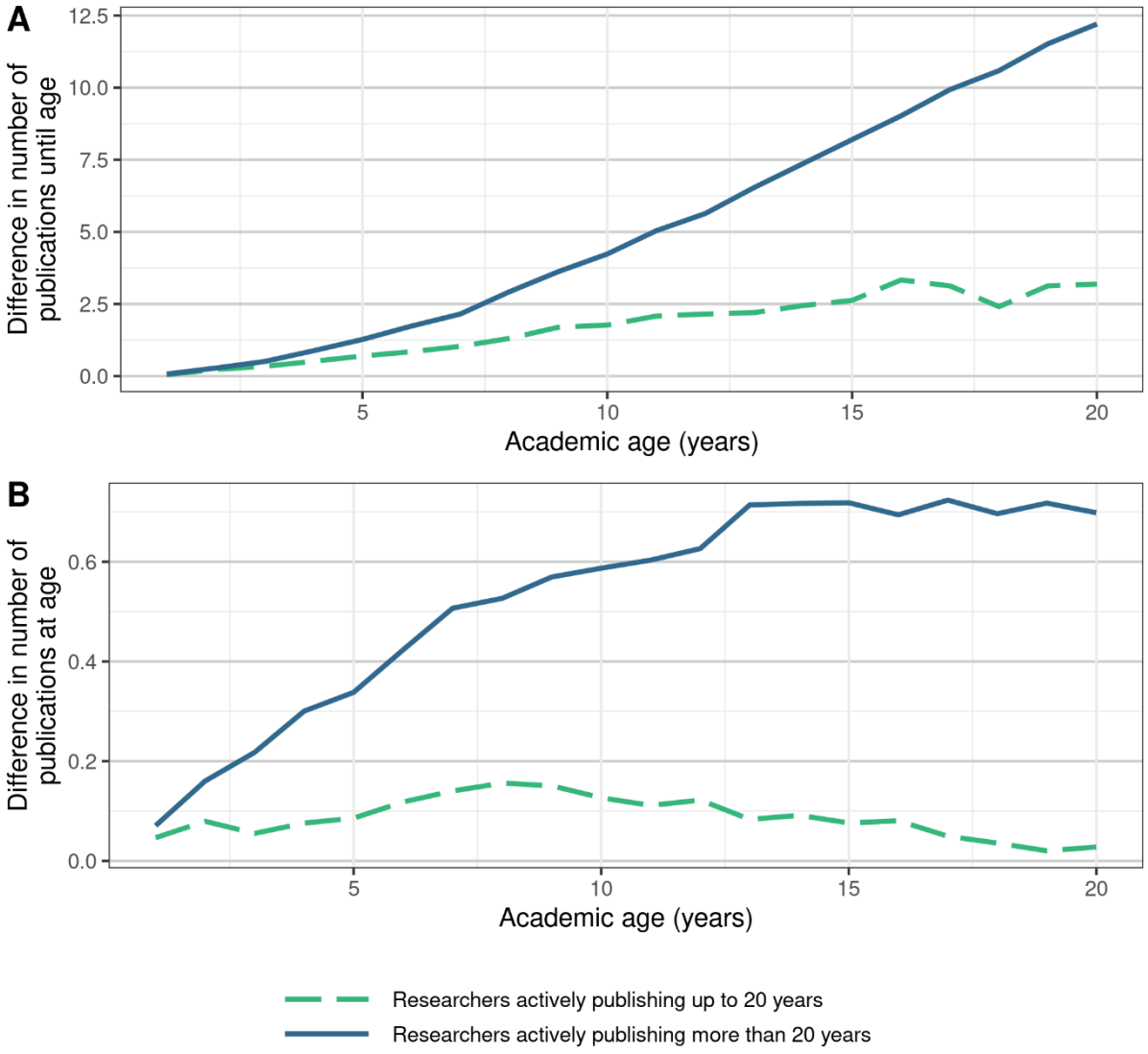


Figure 5-3. Differences in productivity over academic age.
Productivity was measured in terms of the number of papers until a particular academic age (A) and at a particular academic age (B). Differences were calculated between the arithmetic means of male and female researchers (positive values indicate higher productivity for male researchers). The lines show the differences for the researchers who have published their last paper within 20 years after their first publication (green, dashed) and for researchers who have been actively publishing for more than 20 years (blue, solid).

For the results shown in Figure 5-1 and Figure 5-2, the disciplines in which researchers have published were not controlled for. For the results shown in Figure 5-4, this factor was accounted for by applying the matching approach described in Section 5.3.2. Average treatment effects (i.e., gender differences after controlling for disciplines) are plotted over academic age in the figure. The gender difference in productivity considerably decreases after controlling for the disciplines in which researchers have published. For short careers, it nearly disappears. For long careers, it is also much smaller, although some gender difference remains, as the results indicate a higher productivity for male researchers with long careers than female researchers with long careers, even after matching researchers who have published in similar disciplines.

In contrast to productivity, the results for citation impact do not considerably change compared to the results without matching researchers who have published in similar disciplines. This comes as no surprise as citation impact was measured in terms of Hazen percentiles, which are already field-normalised. The results for journal prestige are also quite similar to those without controlling for the disciplines in which researchers publish. Female researchers have generally published in journals with higher impact than male researchers, this gender difference is larger for researchers with short careers, and the gender difference remains relatively constant over academic age. However, the gender difference in journal prestige is generally slightly larger when controlling for disciplines. Figure 5-6 shows that the results for all three output indicators do not change when further differentiating the career length among researchers with short careers (similar to the results shown in Figure 5-2).

Overall, the results suggest that there are small gender differences in scientific output. On average, a slightly higher average citation impact and journal prestige can be observed for female researchers than for male researchers, which is relatively constant over career length and academic age. Meanwhile, male researchers have a higher average productivity in terms of number of papers published until a certain academic age than female researchers, and this difference increases with academic age. However, the increase is based on a relatively constant gender difference in the number of papers published per year. Whereas gender differences in all three output indicators do not systematically depend on the total career length, controlling for disciplines reduces the gender difference in productivity (but not in citation impact or journal prestige).

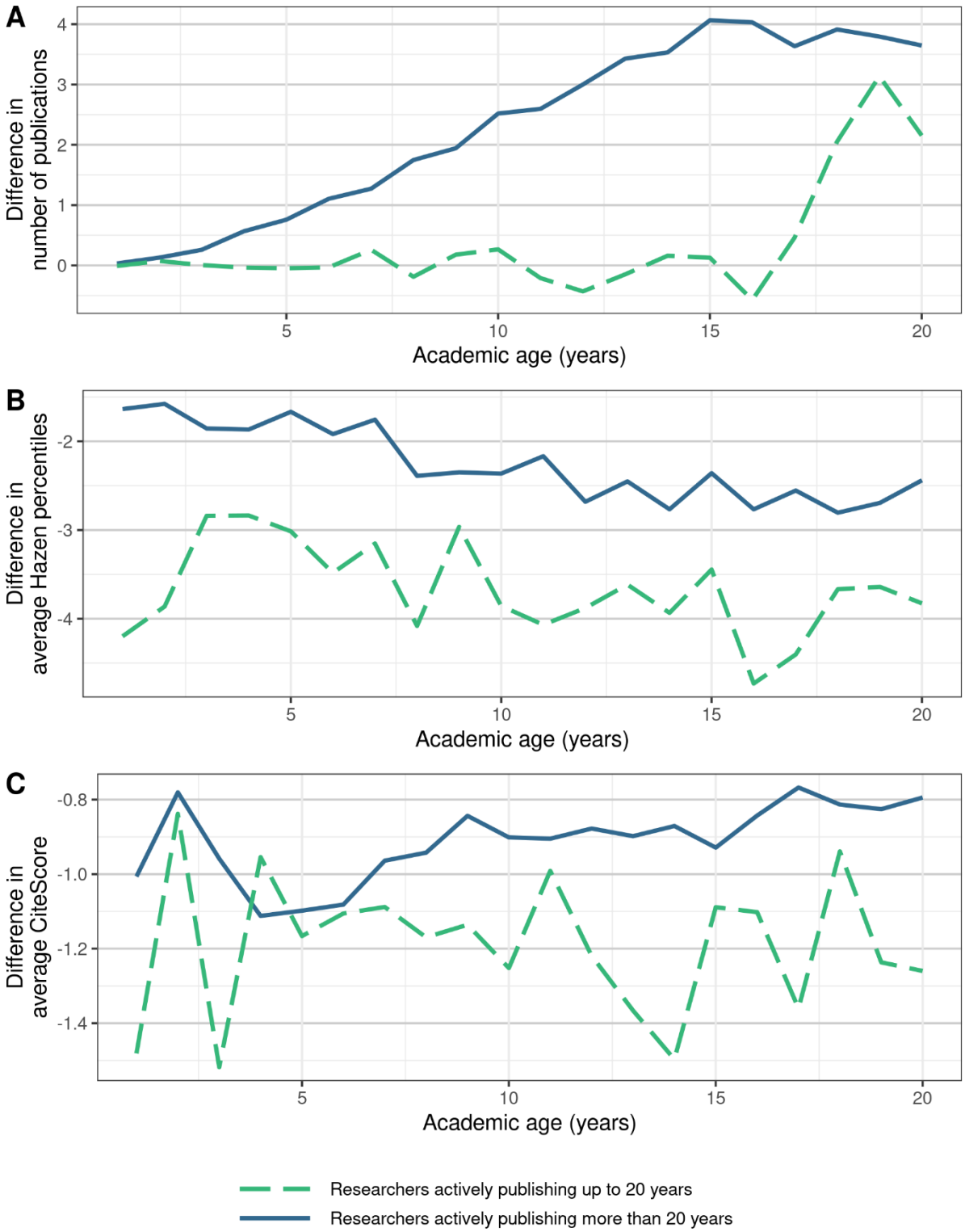


Figure 5-4. Average treatment effects based on matching researchers publishing in similar disciplines.
Scientific output was measured in terms of the total number (A), the average Hazen percentile (B), and the average CiteScore (C) of papers published until a particular academic age. Positive effects indicate higher output for male researchers. The lines show the effects for the researchers who have published their last paper within 20 years after their first publication (green, dashed) and for researchers who have been actively publishing for more than 20 years (blue, solid).

5.5 Discussion

Indicators for measuring scientific output play an important role in researchers' careers and have been discussed as a substantial factor of gender inequalities in science. These indicators are not only relevant for the likelihood of successfully pursuing a scientific career (Berenbaum, 2019; Jungbauer-Gans & Gross, 2013; Thelwall, 2020) but are also regarded as markers of scientific success in themselves. However, it is important to have a good picture of gender differences in scientific output in the first place. Due to different methodological challenges, existing studies are often unable to accurately assess gender differences that are relevant for the likelihood of succeeding in a scientific career. The current study addresses these issues by measuring scientific output at the individual level and controlling for the disciplines in which researchers have published as well as career length. The results consistently demonstrate that, on average, female researchers have a higher citation impact and publish in more prestigious journals than their male peers. By contrast, male researchers tend to publish more papers than female researchers. This productivity difference decreases after controlling for disciplines and only some difference remains for researchers with long careers.

To control for disciplines at the individual level, researchers who have published in similar disciplines were matched. This approach provides an alternative to assigning each researcher to one or a few disciplines, which would align with ordinary field classification systems that assign an entity (journal, paper or researcher) to one or a few disciplines (Waltman & van Eck, 2019). The decrease in gender differences in productivity after controlling for disciplines suggests that gender-specific segregation into disciplines (partly) causes gender differences that can be observed when not adequately considering the disciplines in which researchers publish.

While the methodological approach used in this study allows a better comparison of female and male researchers that are active in similar contexts than many other studies, the analyses focus on a specific perspective on gender differences in science. To develop a comprehensive picture of gender differences in science, other perspectives must also be considered. For example, other dimensions of scientific output could be analysed, such as the novelty (Uddin & Khan, 2016; Uzzi et al., 2013; Wang et al., 2017), disruptiveness (Bornmann et al., 2020; Wu et al., 2019) or breadth (Bu et al., 2021) of a paper's citation impact. Gender differences in science could also be analysed with regard to contributions to the science system other than a researcher's publications, like "teaching, administrative, industrial, or government related research activities" (Huang et al., 2020, p. 4610).

Furthermore, the empirical analyses in this study cannot identify gender differences before or shortly after entering the science system. Researchers with fewer than three publications in the data were not considered, including those who have never published and likely many researchers with short careers. One hypothesis that could not be tested due to this restriction is that the female researchers included in the analyses (i.e., those who have actively published for at least a few years) may be more selective than the male researchers included in the analyses. Such a

higher selectivity of female researchers may also explain the higher citation impact and journal prestige of female researchers, as female researchers who have actively published for at least a few years may be more motivated or qualified to produce papers with high impact and publish in more prestigious journals than male researchers who have actively published for at least a few years.

The results suggest that gender differences in scientific output do not significantly contribute to the underrepresentation of women in science. Female researchers have a higher citation impact and publish in more prestigious journals than male researchers, which would rather support women to pursue a scientific career. Productivity differences with male researchers publishing more papers than female researchers can only be observed for long careers (after controlling for disciplines), which means that the female researchers in this group were still able to pursue a long scientific career despite the productivity differences. These results do not exclude the possibility that higher scientific output might increase female researchers' chances of having a long scientific career. For example, female researchers may need to have a higher scientific output than male researchers to be selected for a job. However, even though some studies suggest that gender bias may influence hiring decisions (Moss-Racusin Corinne et al., 2012), the overall empirical evidence does not support the argument that double standards in hiring processes affect women's chances to be hired (e.g., Auspurg et al., 2017; Ceci et al., 2014). Thus, the underrepresentation of women in science seems to be driven mainly by factors other than gender differences in scientific output.

Although the results of this study reveal no gender differences in scientific output that are likely to contribute to the underrepresentation of women in science, they suggest that the science system loses disproportionately much potential among female researchers. This highlights the importance of fixing the leaky pipeline for women in science to retain the most talented researchers in science, regardless of their gender. The results also have implications for the use of bibliometric indicators in evaluation contexts, as the choice of bibliometric indicators may differentially affect the evaluation of female and male researchers. For example, an average female researcher has a slightly lower level of productivity, but a higher citation impact than an average male researcher who publishes in similar disciplines. If only productivity is used in an evaluation, the female researcher would be evaluated worse as the male researcher. Using citation impact instead, the female researcher would be evaluated better than the male researcher. To mitigate this issue, bibliometric indicators should be carefully selected for evaluation tasks, and several indicators should be used.

5.6 Appendix

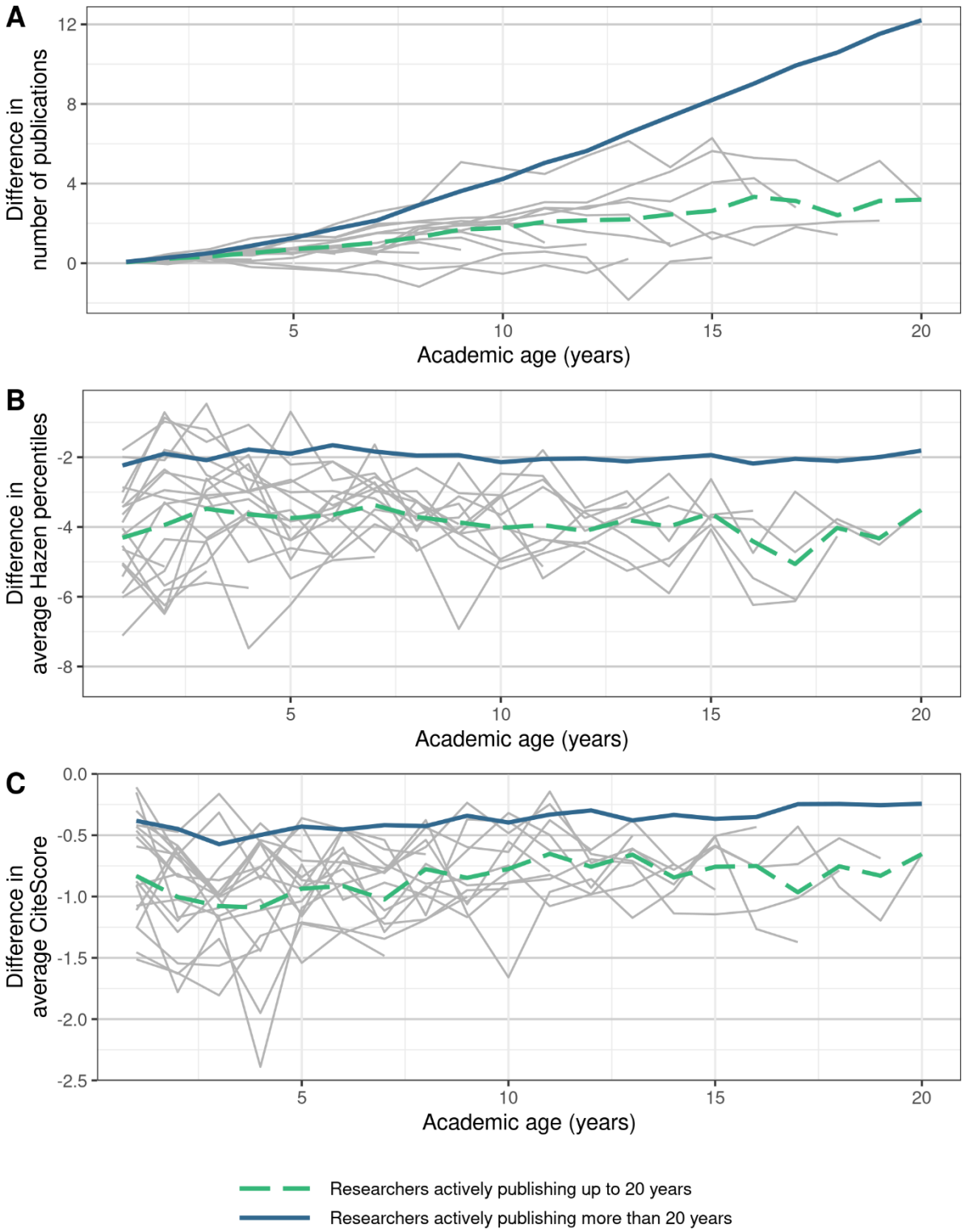


Figure 5-5. Differences in scientific output over academic age.
Scientific output was measured in terms of the total number (A), the average Hazen percentile (B), and the average CiteScore (C) of papers published until a particular academic age. Differences were calculated between the arithmetic means of male and female researchers (positive values indicate higher output for male researchers). The grey lines show the differences separately for different career lengths. The coloured line shows the differences for the pooled sample of researchers across all career lengths.

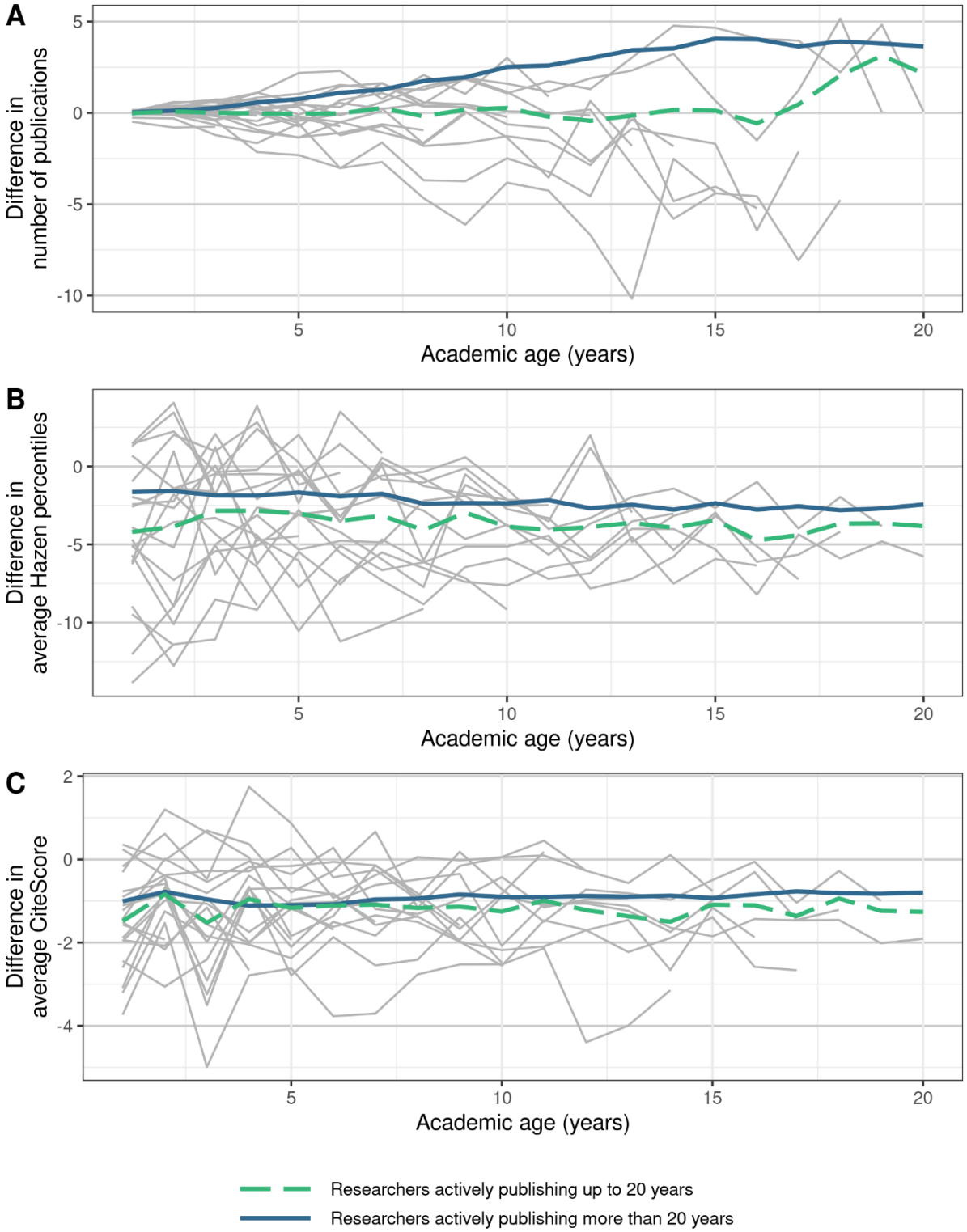


Figure 5-6. Average treatment effects based on matching researchers publishing in similar disciplines.
Scientific output was measured in terms of the total number (A), the average Hazen percentile (B), and the average CiteScore (C) of papers published until a particular academic age. Positive effects indicate higher output for male researchers. The grey lines show the effects separately for different career lengths. The coloured line shows the effects for the pooled sample of researchers across all career lengths.

References

Aaltojärvi, I., Arminen, I., Auranen, O., & Pasanen, H.-M. (2008). Scientific productivity, web visibility and citation patterns in sixteen Nordic sociology departments. *Acta Sociologica*, *51*(1), 5-22. https://doi.org/10.1177/0001699307086815

Akbaritabar, A., & Squazzoni, F. (2020). Gender patterns of publication in top sociological journals. *Science, Technology, & Human Values*, *46*(3), 555-576. https://doi.org/10.1177/0162243920941588

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, *117*(2), 705-720. https://doi.org/10.1007/s11192-018-2895-3

Andersen, J. P., Schneider, J. W., Jagsi, R., & Nielsen, M. W. (2019). Gender variations in citation distributions in medicine are very small and due to self-citation and journal prestige. *eLife*, *8*, e45374. https://doi.org/10.7554/eLife.45374

Auspurg, K., Hinz, T., & Schneck, A. (2017). Berufungsverfahren als Turniere: Berufungschancen von Wissenschaftlerinnen und Wissenschaftlern. *Zeitschrift für Soziologie*, *46*(4), 283-302. https://doi.org/10.1515/zfsoz-2017-1016

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377-386. https://doi.org/10.1162/qss_a_00019

Bendels, M. H. K., Müller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLoS One*, *13*(1), e0189136. https://doi.org/10.1371/journal.pone.0189136

Berenbaum, M. R. (2019). Impact factor impacts on early-career scientist careers. *Proceedings of the National Academy of Sciences*, *116*(34), 16659-16662. https://doi.org/10.1073/pnas.1911911116

Bittmann, F., Tekles, A., & Bornmann, L. (2021). Applied usage and performance of statistical matching in bibliometrics: The comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example. *Quantitative Science Studies*, *2*(4), 1246-1270. https://doi.org/10.1162/qss_a_00158

Boekhout, H., van der Weijden, I., & Waltman, L. (2021). Gender differences in scientific careers. A large-scale bibliometric analysis. In W. Glänzel, S. Heeffer, P.-S. Chi, & R. Rousseau (Eds.), *Proceedings of the 18th International Conference on Scientometrics & Informetrics* (pp. 145-156). ISSI.

Bornmann, L. (2019). Bibliometric indicators. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. Williams (Eds.), *SAGE Research Methods Foundations*. Sage. https://doi.org/10.4135/9781526421036825851

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259. https://doi.org/10.1162/qss_a_00068

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics*, *124*(2), 1457-1478. https://doi.org/10.1007/s11192-020-03512-7

Breuning, M., & Sanders, K. (2007). Gender and journal authorship in eight prestigious political science journals. *PS: Political Science & Politics*, *40*(2), 347-351. https://doi.org/10.1017/S1049096507070564

Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. *Quantitative Science Studies*, *2*(1), 155-183. https://doi.org/10.1162/qss_a_00109

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75-141. https://www.ncbi.nlm.nih.gov/pubmed/26172066

Chatterjee, P., & Werner, R. M. (2021). Gender disparity in citations in high-impact journal articles. *JAMA Network Open*, *4*(7), e2114509-e2114509. https://doi.org/10.1001/jamanetworkopen.2021.14509

de Kleijn, M., Jayabalasingham, B., Falk-Krzesinski, H. J., Collins, T., Kuiper-Hoyng, L., Cingolani, I., Zhang, J., Roberge, G., Deakin, G., Goodall, A., Whittington, K. B., Berghmans, S., Huggett, S., & Tobin, S. (2020). *The researcher journey through a gender lens: An examination of research participation, career progression and perceptions across the globe*. Retrieved 20 November 2022 from www.elsevier.com/gender-report

Duch, J., Zeng, X. H., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., & Nunes Amaral, L. A. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One*, *7*(12), 1-11. https://doi.org/10.1371/journal.pone.0051332

Ebadi, A., & Schiffauerova, A. (2016). Gender Differences in Research Output, Funding and Collaboration. *International Journal of Humanities and Social Sciences*, *112*, 1370-1375. https://publications.waset.org/pdf/10004671

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, *35*(1), 131-150. https://doi.org/10.1177/0306312705046630

Halevi, G. (2019). Bibliometric studies on gender disparities in science. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 563-580). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_9

Hazen, A. (1914). Storage to be provided in impounding municipal water supply. *Transactions of the American Society of Civil Engineers*, *77*(1), 1539-1640. https://doi.org/10.1061/taceat.0002563

Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Gender bias in scholarly peer review. *eLife*, *6*, e21718. https://doi.org/10.7554/eLife.21718

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*, 429-431. https://doi.org/10.1038/520429a

Holman, L., Stuart-Fox, D., & Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented? *PLoS Biol*, *16*(4), e2004956. https://doi.org/10.1371/journal.pbio.2004956

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609-4616. https://doi.org/10.1073/pnas.1914221117

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, *21*(03n04), 1750011. https://doi.org/10.1142/s0219525917500114

Joanis, S. T., & Patil, V. H. (2022). First-author gender differentials in business journal publishing: Top journals versus the rest. *Scientometrics*, *127*(2), 733-761. https://doi.org/10.1007/s11192-021-04235-z

Jungbauer-Gans, M., & Gross, C. (2013). Determinants of success in university careers: Findings from the german academic labor market. *Zeitschrift für Soziologie*, *42*(1), 74-92. https://doi.org/doi:10.1515/zfsoz-2013-0106

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics*, *103*(3), 1061-1071. https://doi.org/10.1007/s11192-015-1580-z

Kern-Goldberger, A. R., James, R., Berghella, V., & Miller, E. S. (2022). The impact of double-blind peer review on gender bias in scientific publishing: A systematic review. *American Journal of Obstetrics and Gynecology*. https://doi.org/https://doi.org/10.1016/j.ajog.2022.01.030

Larivière, V., Ni, C., Gingras, Y., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211-213. https://doi.org/10.1038/504211a

Larivière , V., & Sugimoto, C. (2017). The end of gender disparities in science? If only it were true... Retrieved 20 November 2022 from https://www.cwts.nl/blog?article=n-q2z294

Lynn, F. B., Noonan, M. C., Sauder, M., & Andersson, M. A. (2019). A rare case of gender parity in academia. *Social Forces*, *98*(2), 518-547.

Mauleón, E., & Bordons, M. (2006). Productivity, impact and publication habits by gender in the area of Materials Science. *Scientometrics*, *66*(1), 199-218. https://doi.org/10.1007/s11192-006-0014-3

Mayer, S. J., & Rathmann, J. M. K. (2018). How does research productivity relate to gender? Analyzing gender differences for multiple publication dimensions. *Scientometrics*, *117*(3), 1663-1693. https://doi.org/10.1007/s11192-018-2933-1

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife*, *8*, e47338. https://doi.org/10.7554/eLife.47338

Moss-Racusin Corinne, A., Dovidio John, F., Brescoll Victoria, L., Graham Mark, J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474-16479. https://doi.org/10.1073/pnas.1211286109

Paik, A. M., Mady, L. J., Villanueva, N. L., Goljo, E., Svider, P. F., Ciminello, F., & Eloy, J. A. (2014). Research productivity and gender disparities: A look at academic plastic surgery. *Journal of Surgical Education*, *71*(4), 593-600. https://doi.org/https://doi.org/10.1016/j.jsurg.2014.01.010

Raj, A., Carr, P., Kaplan, S. E., Terrin, N., Breeze, J. L., & Freund, K. M. (2016). Longitudinal analysis of gender differences in academic productivity among medical faculty across 24 medical schools in the United States. *Acad Med.*, *91*(8). https://doi.org/10.1097/ACM.0000000000001251

Reijnhoudt, L., Costas, R., Noyons, E., Börner, K., & Scharnhorst, A. (2014). 'Seed + expand': a general methodology for detecting publication oeuvres of individual researchers. *Scientometrics*, *101*(2), 1403-1417. https://doi.org/10.1007/s11192-014-1256-0

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P., & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances*, *7*(2), eabd0299. https://doi.org/10.1126/sciadv.abd0299

Studer, C. (2012). GitHub repository cstuder/genderReader. (March 20, 2019), Retrieved from https://github.com/cstuder/genderReader. https://github.com/cstuder/genderReader

Teele, D. L., & Thelen, K. (2017). Gender in the journals: Publication patterns in political science. *PS: Political Science & Politics*, *50*(2), 433-447. https://doi.org/10.1017/S1049096516002985

Tekles, A., Auspurg, K., & Bornmann, L. (2022). Same-gender citations do not indicate a substantial gender homophily bias. *PLoS One*, *17*(9), e0274810. https://doi.org/10.1371/journal.pone.0274810

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches1. *Quantitative Science Studies*, *1*(4), 1510-1528. https://doi.org/10.1162/qss_a_00081

Thelwall, M. (2018). Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries. *Journal of Informetrics*, *12*(4), 1031-1041. https://doi.org/https://doi.org/10.1016/j.joi.2018.08.005

Thelwall, M. (2020). Gender differences in citation impact for 27 fields and six English-speaking countries 1996–2014. *Quantitative Science Studies*, *1*(2), 599-617. https://doi.org/10.1162/qss_a_00038

Thelwall, M., Abdoli, M., Lebiedziewicz, A., & Bailey, C. (2020). Gender disparities in UK research publishing: Differences between fields, methods and topics. *El profesional de la información*, e290415. https://doi.org/10.3145/epi.2020.jul.15

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, *10*(4), 1166-1177. https://doi.org/https://doi.org/10.1016/j.joi.2016.10.004

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472. https://doi.org/10.1126/science.1240474

van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2012). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, *93*(3), 857-868. https://doi.org/10.1007/s11192-012-0712-y

van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: a longitudinal case study. *Scientometrics*, *106*, 143-162. https://www.ncbi.nlm.nih.gov/pubmed/26798162

van den Besselaar, P., & Sandström, U. (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLoS One*, *12*(8), e0183301. https://doi.org/10.1371/journal.pone.0183301

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20-41. https://doi.org/10.1162/qss_a_00112

Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 281-300). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_11

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851-872. https://doi.org/10.1007/s11192-012-0775-9

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416-1436. https://doi.org/https://doi.org/10.1016/j.respol.2017.06.006

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS One*, *8*(7), e66212. https://doi.org/10.1371/journal.pone.0066212

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, *393*(10171), 531-540. https://doi.org/https://doi.org/10.1016/S0140-6736(18)32611-4

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-382. https://doi.org/10.1038/s41586-019-0941-9

Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., Woodruff, T. K., & Amaral, L. A. N. (2016). Differences in collaboration patterns across discipline, career stage, and gender. *PLOS Biology*, *14*(11), e1002573. https://doi.org/10.1371/journal.pbio.1002573

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics*, *126*(11), 8861-8886. https://doi.org/10.1007/s11192-021-04171-y