Should citations be field-normalized in evaluative bibliometrics?

An empirical analysis based on propensity score matching

Lutz Bornmann*, Robin Haunschild** & Rüdiger Mutz***

*Division for Science and Innovation Studies
Administrative Headquarters of the Max Planck Society
Hofgartenstr. 8,
80539 Munich, Germany.
E-mail: bornmann@gv.mpg.de

**Max Planck Institute for Solid State Research
Heisenbergstraße 1,
70569 Stuttgart, Germany.
Email: r.haunschild@fkf.mpg.de

***Professorship for Social Psychology and Research on Higher Education
ETH Zurich
Andreasstrasse 15 (AND 2 80),
8050 Zurich, Switzerland.
Email: ruediger.mutz@gess.ethz.ch
Abstract

Field normalization of citations is bibliometric standard. Despite the observed impact differences between fields, the question remains how strong fields influence the citation impact of papers. We wondered whether field differences can be traced back to other factors (than fields) possibly influencing citations (FICs) or are “pure” field effects. We considered several FICs such as number of pages and number of co-authors. Using the propensity score matching approach, we investigated with Web of Science data whether papers with the same or similar attributes (e.g., number of co-authors and number of pages) received different citations when they were assigned to different fields. In a diagnostic step of our statistical analyses, we considered propensity scores as covariates in regression analyses to examine whether the differences between the fields vanish. The results revealed that the differences did not completely vanish, but were significantly reduced. We received similar results when we calculated mean value differences of the fields after inverse-probability of treatment weighting representing the causal or unconfounded field effects on citations. One possible interpretation of our results is that field-differences in citations exist which are relatively independent of other FICs. Another possible interpretation is that field-differences could vanish in principle.

Key words

bibliometrics; field-normalization; propensity score matching
1 Introduction

In the 1980s, bibliometrics changed from a field comprising some active professionals (e.g., Robert K. Merton, Jonathan R. Cole, and Stephen Cole) to a professional field including more and more bibliometric specialists (e.g., Wolfgang Glänzel and Anthony F.J. van Raan). At the same time, bibliometric methods have increasingly been used for larger entities such as research groups, departments, institutes, universities, and countries (Moed, 2017). The broadening of bibliometrics to larger entities also meant that bibliometricians saw the need to tackle the problem of field-specific citation rates in their analyses. This need was based on the observation that fields have (very) different citation rates: for example, the average citation rate in biology is much higher than that in mathematics and the time to get recognized and cited differs in both fields (Wang, 2013).

Today, field-specific citation rates are targeted by applying field-normalized indicators in citation analyses (Waltman, 2016). The citation impact of a focal paper is standardized based on the citation impact of papers in a reference set. Field-normalization is a standard concept in professional bibliometrics (Leydesdorff, Wouters, & Bornmann, 2016) which is not only widely used (Purkayasthaa, Palmaroa, Falk-Krzesinskib, & Baas, 2019), but also recommended as one of ten guiding principles for evaluation studies in the Leiden manifesto (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Wilsdon et al., 2015). Thus, field-normalization is a sign of professional bibliometrics and one of the few concepts, the necessity of which scarcely any bibliometrician would doubt. Much efforts in bibliometrics research is spent on various technical improvements in the calculation of field-normalized indicators (Wilsdon et al., 2015).

A list of possible reasons for field-specific differences in citation impact has been published by Waltman and van Eck (2013b): “Each field has its own publication, citation, and authorship practices, making it difficult to ensure the fairness of between-field comparisons.
In some fields, researchers tend to publish a lot, often as part of larger collaborative teams. In other fields, collaboration takes place only at relatively small scales, usually involving no more than a few researchers, and the average publication output per researcher is significantly lower. Also, in some fields, publications tend to have long reference lists, with many references to recent work. In other fields, reference lists may be much shorter, or they may point mainly to older work. In the latter fields, publications on average will receive only a relatively small number of citations, while in the former fields, the average number of citations per publication will be much larger” (p. 833). Similar points are mentioned by Abramo, Cicero, and D’Angelo (2011). According to Wang, Song, and Barabási (2013), “comparison of different papers is confounded by incompatible publication, citation, and/or acknowledgment traditions of different disciplines and journals” (p. 127).

Field-specific impact differences have been published in many bibliometric studies. However, we could not find any study, which has investigated the causal link between field assignment and citation impact of papers. The above mentioned reasons by Waltman and van Eck (2013b) for explaining field-specific citation rates point to possible confounding factors that are specific for both citation impact and field assignment causing a spurious association. For example, one reason for field-specific differences of citation rates are – according to Waltman and van Eck (2013b) – differences in the number of co-authors. One simple reason for the correlation between the number of co-authors and the number of citations might be the increasing number of self-citations with more co-authors. Since both field and number of co-authors might have an influence on citations, the relationship between field and citation impact could be confounded by the number of co-authors. Using a small world example, we would like to demonstrate this possible confounding.

Table 1 shows a small world example, which reveals that the number of authors might cause differences in citation rates which are also reflected in field-specific citation rates. Based on the data in the table, field 1 (m=62.5) has a higher citation rate than field 2
Thus, one could have the (false) impression that field specifics cause different citation rates. However, the view on the number of co-authors demonstrates that it is actually this variable, which determines differences in mean citations: a higher number of co-authors usually leads to a higher number of citation counts. Papers with five co-authors have 71.7 citations on average, and papers without co-authors 1.7 citations. Paper 3 makes the difference: its high citation impact (90 citations) does not correspond to the impact of the corresponding field (m=23.75 for field 2), but to the impact, which can be expected for five authors (m=71.7).

Table 1. Small world example, which demonstrates that citations might be more dependent on the number of co-authors than on the field

| Paper    | Number of co-authors | Field | Citations |
|----------|-----------------------|-------|-----------|
| Paper 1  | 5                     | 1     | 50        |
| Paper 2  | 5                     | 1     | 75        |
| Paper 3  | 5                     | 2     | 90        |
| Paper 4  | 1                     | 2     | 0         |
| Paper 5  | 1                     | 2     | 1         |
| Paper 6  | 1                     | 2     | 4         |

Many empirical studies in the field of bibliometrics have shown that not only the number of authors, but also many other factors are possibly related to citation impact differences between papers – besides their scientific quality. As factors potentially influencing citations (FICs) the number of pages, the reputation of the publishing journal, and many other factors have been identified (see overviews of FICs in Tahamtan & Bornmann, 2018, in press; Tahamtan, Safipour Afshar, & Ahamd zadeh, 2016). Having these many FICs in mind, it becomes suspicious that visible field effects in citation rates can be traced back to underlying factors, which (at least partially) cause these differences. Since we know that different numbers of co-authors are related to differences in citation rates (because of, e.g., a self-citation effect), it might be the case that fields have different citation rates, because the
average number of authors per field is different. Thus, the field effect is actually an author effect. The same might be true for any other FIC besides the number of co-authors (e.g., the number of pages).

In this study, we address the question of whether differences in citation rates can be traced back to field variations or whether the differences are the results of other FICs than fields (e.g., number of co-authors or number of pages). We apply propensity score matching – an approach which has been introduced for measuring causal links – to test whether papers with the same characteristics (e.g., number of co-authors and number of pages) received different citation impact when they have been assigned to different fields. It is the aim of the present study to identify the “true” field effect: is there a causal field effect besides usual characteristics of fields (e.g., number of co-authors or number of cited references) which makes it necessary to field-normalize – beyond these characteristics. The fruitfulness of propensity score matching for bibliometric studies has been demonstrated by Mutz and Daniel (2012a), Mutz, Wolbring, and Daniel (2017) and Farys and Wolbring (2017). The results of the present study may have far-reaching consequences for the approaches used in advanced bibliometrics: either the necessity of field-normalization is justified (and the current field-normalized indicators remain standard approaches) or it is questioned. In the latter case, other indicators for citation impact normalization should be developed – possibly considering other FICs than fields.

2 Field-differences in citation rates and normalization of citation impact in bibliometrics: a literature overview

The Leiden Manifesto for research metrics includes ten principles to guide research evaluation (Hicks et al., 2015). Principle six states that normalized citation impact indicators are required (instead of raw citation counts) because publication and citation practices vary by field (see also Wilsdon et al., 2015). According to Ioannidis, Boyack, and Wouters (2016)
“the basic premise of normalization is that not all citations are equal”. Later, these authors write that “one wants to correct for imbalance of citation opportunity” (Ioannidis et al., 2016). Crespo, Li, and Ruiz-Castillo (2013) list the following characteristics demonstrating differences in publication and citation practices: “(i) size, measured by the number of publications in the periodical literature; (ii) the average number of authors per paper; (iii) the average paper length; (iv) the average number of papers per author over a given period of time; (v) the theoretical or experimental mix that characterizes each discipline; (vi) the average number of references per paper; (vii) the proportion of references that are made to other articles in the periodical literature; (viii) the percentage of internationally co-authored papers, or (ix) the speed at which the citation process evolves”.

Field-normalized indicators have been developed to enable comparisons between the citation impact of publications from different fields (Mingers & Leydesdorff, 2015; Waltman, 2016). These comparisons are necessary if units are evaluated publishing in different fields (e.g., universities or countries). One of the most important field-normalized indicators is the mean-normalized citation score which relates the citation impact of a unit to a worldwide, field-specific reference value (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Percentile-based indicators focus on the number or proportion of publications (of a unit) that belong to the top 10% or the top 1% of their field (Wilsdon et al., 2015). Waltman et al. (2012) regard the top 10% indicator “as the most important impact indicator in the Leiden Ranking” (p. 2425). Another frequently applied approach for normalizing citation data was proposed in the 1980s – the Characteristic Scores and Scales (CSS) method (Glänzel & Schubert, 1988). Here, the publications in reference sets are classified as follows:

“characteristic scores are obtained from iteratively truncating a distribution according to conditional mean values from the low end up to the high end. In particular, the scores \( b_k \) (\( k > 0 \)) are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or no new scores are obtained”
(Glänzel, 2013, p. 111). Mutz and Daniel (2019) have recently suggested a statistical approach to model raw citations but considering reference values in the statistical model.

One of the most frequently discussed topics in bibliometrics related to field-normalization concerns the delineation of fields. The boundaries of fields can be defined and implemented in field-normalization in various ways. It seems impossible to verify satisfactorily that one solution is better suited than another for field-normalization. Most bibliometricians use multi-disciplinary classification systems as the basis for building reference sets, which have a broad coverage of publications from various disciplines. Among these systems, the most popular are the subject categories based on sets of journals provided by Clarivate Analytics in the Web of Science (WoS, about 250 journal-based subject categories) or by Elsevier in Scopus (Wang & Waltman, 2016; Wouters et al., 2015) spanning nearly all fields (Sugimoto & Weingart, 2015). The use of journal sets for defining fields seems obvious because fields become mature when they introduce their first professional journal (Sugimoto & Weingart, 2015). However, the results by Wang and Waltman (2016) suggest that “WoS and especially Scopus tend to be too lenient in assigning journals to categories. A significant share of the journals in both databases, but especially in Scopus, seem to have assignments to too many categories” (p. 359).

If journal sets are used for field categorization, each paper belongs to the subject category (one or more) of its publishing journal. One of the problems with this field classification approach is multi-disciplinary journals (e.g. Nature or Science); the papers in these journals cannot be assigned to meaningful subject categories based on journal classification (Kronman, Gunnarsson, & Karlsson, 2010). The journal classification system is also stretched to its limits in the case of emerging or interdisciplinary fields, because the papers in these fields are usually published in a wide range of different journals (Strotmann & Zhao, 2010). Another problem might be the fact that many journal categories are too broad
(Haddow & Noyons, 2013). They seem to cover multiple fields (each with their own citation practices) in a single category (van Eck, Waltman, van Raan, Klautz, & Peul, 2013).

Instead of journal-based field classification systems, publication-based field classification systems can be used for normalizing citations which might offer a more fine-grained representation of fields than the journal-based system (Waltman & van Eck, 2019). As an alternative to journal sets, Waltman and van Eck (2012) introduced a method for algorithmically constructing classification systems (ACCS). The method also leads to a multidisciplinary classification system, which is based on direct citation relations between single papers. This method already plays an important role in scientometrics, because the measuring of field-normalized citation impact in the Leiden Ranking is based on this approach (see https://www.leidenranking.com/information/fields). The results of Klavans and Boyack (2017) indicate that algorithmically constructed classification systems might be more accurate than systems based on other methods: “Using the assumption that higher concentrations of references denote more accurate clusters, direct citation thus provides a more accurate representation of the taxonomy of scientific and technical knowledge than either bibliographic coupling or co-citation” (p. 984).

However, Leydesdorff and Milojević (2015) criticize the approach of algorithmically constructed classifications systems: “because these ‘fields’ are algorithmic artifacts, they cannot easily be named (as against numbered), and therefore cannot be validated. Furthermore, a paper has to be cited or contain references in order to be classified, since the approach is based on direct citation relations” (p. 201). In a case study, Haunschild, Schier, Marx, and Bornmann (2018) concluded: “[o]ne possible interpretation of our results is that the cluster algorithm used to construct ACCS is not able to distinguish properly between scientific fields” (p. 445). Haunschild, Marx, French, and Bornmann (2018) compared field-and time-normalized indicators using three different classification schemes: (i) WoS journal sets, (ii) the ACCS proposed by Waltman and van Eck (2012), and (iii) an intellectually-based
classification system by experts from the Chemical Abstracts Service (CAS) for 2,690,143 papers in the field of chemistry. They found that the normalized citation scores agree better when WoS journal sets and CAS classifications are used than when either of them is compared with normalized citation scores based on the ACCS proposed by Waltman and van Eck (2012). Remarks on other multi-disciplinary classification systems (which are not very popular in bibliometrics) can be found in Wang and Waltman (2016).

For field-normalization in a specific discipline and its related disciplines, mono-disciplinary classification systems can be used instead of multi-disciplinary systems. Mono-disciplinary systems are usually based on intellectual classifications of papers (Bornmann, Marx, & Barth, 2013). This means that human indexers classify the publications by using a detailed and controlled vocabulary. Either experts in disciplines classify papers, as with papers in the Chemical Abstracts (CA) database (Bornmann & Daniel, 2008; Bornmann, Mutz, Neuhaus, & Daniel, 2008), or the authors of papers themselves undertake this exercise, as in economics with the Journal of Economic Literature (JEL) classification system (Bornmann & Wohlrabe, 2019). Several classification systems have already been used for field-normalization besides CA and JEL: Medical Subject Headings (MeSH) (Boyack, 2004), Physics and Astronomy Classification Scheme (PACS) (Radicchi & Castellano, 2011), and sections of the American Sociological Association (ASA) (Teplitskiy & Bakanic, 2016).

Another advantage of mono-disciplinary classification systems is that the classification provides a relatively high level of detail (Wang & Waltman, 2016). It is, however, a disadvantage of these systems that they cannot be used for field-normalization exercises referring to a broad spectrum of disciplines.

The different approaches for assigning publications to fields which are used in bibliometrics to normalize citations are only vaguely connected to criteria for defining fields (Hammarfelt, 2018). Bibliometricians seem to prefer practically realizable approaches neglecting the criteria which Sugimoto and Weingart (2015) denote as (1) cognitive (shared
bodies of content, theories, and methods), (2) social (communities of multiple researchers using specific terminologies or technical languages), and (3) institutional (institutionalization in the form of academic departmentalization). According to Sugimoto and Weingart (2015), “the maturation of a discipline is perhaps best represented by its institutionalization, particularly in the form of academic departmentalization” (p. 780). Andersen (2016) focusses on the cognitive and social dimensions in characterizing fields: “A scientific discipline (or specialty, field or domain) could be understood … as an epistemic unit consisting of a set of closely related cognitive resources such as, for example, concepts, models and theories, and as a social unit consisting of highly similar experts who were employing and at the same time developing their shared cognitive resources” (p. 2). A field can be defined as an “invisible college” (Ziman, 1996, p. 69) sharing a particular research tradition, having a “great man” (Sugimoto & Weingart, 2015, p. 782) in its history, and having access to specific governmental funding (e.g., specific programs for nanotechnology).

The basic reason for conducting field-normalization in bibliometrics is that impact should be measured without confounding factors: “ideally, one wants citation indicators to measure impact in a monotonic fashion: the higher the metric, the ‘better’ the paper” (Ioannidis et al., 2016). In the definition of field-normalized indicators, not only the field of papers is considered, but also their publication years and document type. According to the document type, many studies published hitherto have found that “reviews” usually receive more citations than “articles” – both are the most important document types in bibliometric analyses (e.g., Lundberg, 2007). Besides field, time, and document type, many other FICs have been identified (e.g., number of authors, length of paper title, and number of cited references). Comprehensive overviews of these factors can be found in Tahamtan and Bornmann (2018), Hanel and Haase (2017), Onodera and Yoshikane (2014), and Didegah and Thelwall (2013).
The identification of FICs – as important influencing factors on citations besides cognitive influence – is rooted in the social-constructivist side of citation theories. This side stresses the rhetorical functions of citations and the impossibility to standardize citation behavior across scientists (Riviera, 2013): authors cite the same paper for different reasons and citations have different functions. For instance, Gilbert (1977) introduced the idea that referencing is an aid to persuasion of readers. The micro-sociological school, which represents the social-constructivist theory tradition, contends that “scientists may have complex citation motives that have not yet been clearly understood, and therefore, they cannot be satisfactorily described unidimensionally” (Liu, 1993, p. 13). Although all cited and citing works are related by the same link in citation indexes, the authors’ motives and citation behavior can be different (Tahamtan & Bornmann, in press).

The other side of citation theory is reserved by the normative theory of citation (Merton, 1973) which assumes that authors follow uniform standards in citing publications. According to this theory, citations can be used for evaluative purposes, since citers give credits to cited publications (or cited scientists) which have influenced their research. According to Merton (1988) “the reference serves both instrumental and symbolic functions in the transmission and enlargement of knowledge. Instrumentally, it tells us of work we may not have known before, some of which may hold further interest for us; symbolically, it registers in the enduring archives the intellectual property of the acknowledged source by providing a pellet of peer recognition of the knowledge claim, accepted or expressly rejected, that was made in that source” (p. 622).

Basically, both theories of the citation process try “to explain why author \( x \) cite[s] article \( a \) at time \( t \)” (Sandström, 2014, p. 59). Research suggests that both theories have their place in explaining citation behavior. The study by Judge, Cable, Colbert, and Rynes (2007) suggests that “universalistic, particularistic, and mixed universalistic-particularistic characteristics all play significant roles in the extent to which research articles in the field of
management are cited” (p. 500). The study by Yu, Yu, Li, and Wang (2014) demonstrates that a paper’s citation impact can be predicted based on several factors with a potential influence on citations: “External features of a paper, features of authors, features of the journal of publication, and features of citations are all involved in constructing a paper’s feature space through the mathematical description method … With relative accuracy, we can predict paper citation impact after the first 5 years of publication in this subject” (p. 1250). On one side, the results of the study question the normative theory of citation that the quality of papers is the main factor driving citations. On the other side, the study demonstrates the necessity to consider many FICs if one is intended to explain citation impact of papers.

3 Methods

3.1 Dataset used

We used the papers of the WoS from our in-house database – derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) provided by Clarivate Analytics (formerly the IP and Science business of Thomson Reuters). Citations were counted until the end of 2017.

Since it was not possible to base the statistical analyses on the complete in-house database, we selected a stratified sample covering a broad range of subject categories and as many as possible citation impact differences between the subject categories. We exported the meta-data of all English papers and their citation counts of the document type “article” which were published between 2000 and 2005 and assigned to at least one of the following ten WoS subject categories: (1) Biodiversity conservation, (2) Computer science, Artificial intelligence, (3) Communication, (4) Engineering, petroleum, (5) Family studies, (6) Geriatrics & gerontology, (7) Immunology, (8) Physics, particles & fields, (9) Rehabilitation, and (10) Spectroscopy. We kept only those publications for which we had sufficient information about relevant meta-data, e.g., number of pages, number of co-authors, authors’
countries, and Journal Impact Factor (JIF, the average citation impact of the papers published in a journal within one year) are not missing.

The set of these WoS subject categories was selected as follows: The analyzed WoS subject categories should not have too low average citation counts, but they should differ as much as possible in their number of average citation rates. We selected each twentieth WoS subject category in the database ordered by their average citation rate in 2005 (starting with that category having the highest average citation rate) from those fulfilling the following criteria: (1) more than 100 papers of the document type “article” were published in 2005, and (2) only papers which belong to one of the following OECD categories\(^1\) are considered: Natural sciences, Engineering and technology, Medical and Health sciences, and Social sciences.

In total, 290,310 articles were included in the analysis.

3.2 Variables included

Table 2 shows the variables, which we included in this study. The variable “total citation counts” is the metric, which we would like to explain in the statistical analyses. As FICs, we included 11 variables (9 factors) in the statistical analyses. We tried to consider as many FICs as possible in our study. The overview of Tahamtan et al. (2016) revealed “28 factors, influencing the frequency of citations … which were classified into three categories: ‘Paper related factors’, ‘Author related factors’ and ‘Journal related factors’” (p. 1198). However, we could not include in the present study all the many FICs proposed (investigated) in the past. The most important reason for the reduction in this study was that many FICs must be preprocessed manually (human-based) and/or are in need of an analysis of the full-text of publications (e.g., the number of differential equations, the academic titles of the authors or the specific design of a study; see Antonakis, Bastardoz, Liu, & Schriesheim, 2014;

\(^1\) See: http://help.prod-incites.com/inCites2Live/filterValuesGroup/researchAreaSchema/oecdCategoryScheme.html
Di Vaio, Waldenström, & Weisdorf, 2012; Robson & Mousquès, 2016). Other FICs cannot be considered, since reliable and valid information is not available for statistical analyses: for example, the phenomenon “obliteration by incorporation” (McCain, 2014) refers to “work that has become so accepted that it is no longer cited” (MacRoberts & MacRoberts, 2010, p. 2).

Table 2. Sample description

| No | Label                        | Mean | Median | Standard deviation | Minimum | Maximum |
|----|------------------------------|------|--------|--------------------|---------|---------|
| 1  | Number of pages              | 9.73 | 8      | 7.33               | 1       | 878     |
| 2  | Number of co-authors         | 5.39 | 3      | 19.37              | 1       | 942     |
| 3  | Number of author addresses   | 3.38 | 3      | 3.27               | 1       | 115     |
| 4  | Number of joined countries   | 1.31 | 1      | 0.89               | 1       | 22      |
| 5  | USA                          | 0.37 | 0      | 0.48               | 0       | 1       |
| 6  | Europe                       | 0.35 | 0      | 0.48               | 0       | 1       |
| 7  | Number of keywords           | 5.41 | 5      | 4.04               | 0       | 10      |
| 8  | Number of title words        | 12.11| 11     | 5.01               | 1       | 61      |
| 8  | Number of cited references   | 28.68| 26     | 19.26              | 0       | 724     |
| 8  | Number of linked cited       | 19.68| 16     | 16.65              | 0       | 593     |
|    | references                   |      |        |                    |         |         |
| 9  | Journal Impact Factor        | 2.39 | 1.56   | 2.53               | 0       | 46.23   |
|    | Total citations              | 29.27| 13     | 84.97              | 0       | 18,348  |

Although it was not possible in this study to consider the complete set of FICs proposed in previous studies, we tried to consider those factors for which many studies have already published meaningful relationships with citation counts. For example, the tabulated overview of FICs published by Hanel and Haase (2017) shows that the JIF and the number of cited references – both FICs were included in this study – have been investigated, see No 9 (JIF) and No 8 (number of cited references) in Table 2. The tabulated overviews by Onodera and Yoshikane (2014) and Didegah and Thelwall (2013) contain the additional important
information for our FICs selection indicating whether the studies having investigated FICs report weak or strong relationships with citations or report contradictory results.

The first FIC in Table 2 is the number of pages or the size of a publication. One might speculate that more extensive papers include more content (empirical results, ideas, conceptual explanations etc.) than papers with only a few pages which might lead to more references to the content. In agreement with this assumption, the results of the early study by Gillmor (1975) show that in general the longer the paper, the more citations it received. The author investigated all paper published in Journal of Atmospheric and Terrestrial Physics between 1967 and 1973. Similar results have been published by Leimu and Koricheva (2005) based on ecological papers and by Stanek (2008) based on papers published in major astronomy journals. According to Falagas, Zarkali, Karageorgopoulos, Bardakas, and Mavros (2013) the correlation between paper length and citation numbers holds true (considering major general medicine journals) “after adjustment for several potentially confounding variables, such as the study design, prospective or retrospective nature of the study, abstract and title word count, number of author-affiliated institutions and number of bibliographic references”.

The second FIC in Table 2 is the number of co-authors (having collaborated for a paper). Many studies have investigated the relationship between this FIC and the number of citations finding a positive correlation (e.g., Beaver, 2004; Benavent-Pérez, Gorraiz, Gumpenberger, & Moya-Anegón, 2012; Lawani, 1986; Tregenza, 2002). For ecology papers, the study by Leimu and Koricheva (2005) reveals that “papers with four or more authors received more citations than did papers with fewer authors” (p. 30). The proposed reasons for the dependence of the number of co-authors and citations are the multi-disciplinarity of papers with many differently skilled co-authors (Wray, 2006) or the benefits of labour division by bringing “complementary talents and expertise to a problem” (Walstad, 2002, p. 20). In addition, “the higher the number of authors, the larger the network of scientists that
might know of one of them and, thus, cite them. Alternatively, the increase in citation rates with the number of authors might be related to an increased frequency of self-citations in the case of multi-authored papers” (Leimu & Koricheva, 2005, p. 30). However, the latter results could only be partly confirmed by Glänzel, Debackere, Thijs, and Schubert (2006). According to Valderas (2007), the longer the author list “the greater the probability of the paper being presented to several conferences is, especially if the team is multidisciplinary” (p. 1496) which might increase the number of citations.

The third FIC in Table 2 is related to the second: **number of author addresses**. One can expect that more co-authors for publications are related to more different author addresses (see Taborsky, 2009). For papers published in nanoscience and nanotechnology journals from 2007 to 2009, Didegah and Thelwall (2013) found a positive relationship between the number of institutions (mentioned on a paper) and the number of citations. The same has been reported for papers published by researchers affiliated with a Spanish university (Iribarren-Maestro, Lascurain-Sanchez, & Sanz-Casado, 2007). Although the number of co-authors and the number of author addresses might be positively correlated, both might not be related to citations in the same way: collaborations of authors between different institutions can lead to higher citation rates than collaborations within the same institutions (Whitfield, 2008). A similar result has been found by Dong, Ma, Tang, and Wang (2018) measuring innovation of research: “we discover striking phenomena where a smaller, yet more diverse team is more likely to generate highly innovative work than a relatively larger team within one institution” (see also Jones, Wuchty, & Uzzi, 2008).

The fourth FIC in Table 2 reflects a similar concept as the second and third FICs do: **number of joined countries**. The above mentioned study by Iribarren-Maestro et al. (2007) did not only find a positive correlation between the number of institutions and the number of citations, but also between the number of countries (indicated by the co-authors’ addresses of a paper) and citation impact. Reasons for an increased citation impact of papers published in
international collaborations might be the greater spreading of published papers and the greater probability of including talented people in a team of co-authors – less talented co-authors might have been found in the same country.

The fifth FIC in Table 2 also focusses on countries, but targets an effect other than the number of countries. We included the information in the statistical analysis, whether an author from the USA or Europe (EU28: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and United Kingdom) published the paper. These binary variables focus on possible national biases in citations (see Bornmann, Stefaner, de Moya Anegon, & Mutz, 2014; Bornmann, Wagner, & Leydesdorff, 2018; Lariviere, Gong, & Sugimoto, 2018). According to the literature overview by Taborsky (2009) “regional affiliations” and “whether the authors are from native English speaking countries” (p. 105) are related factors with citations. The results by Pasterkamp, Rotmans, de Kleijn, and Borst (2007) show that “citation frequency was significantly augmented by nation oriented citation bias. This nation oriented citation behaviour seems to mainly influence the cumulative citation number for papers originating from the countries with a larger research output” (p. 153). Thus, one can expect papers with high citation impact from high research-performing countries. In other words, “international collaboration alone is not important, unless it is with a high impact nation” (Didegah & Thelwall, 2014, p. 174). The results of Gingras and Khelfaoui (2018) reveal that countries especially profit from collaborations with the USA.

The sixth FIC in Table 2 is the number of key words, which are mentioned in a paper. This number might reflect not only the spread of topics a paper deals with, but also its inter-disciplinarity. More key words might indicate that topics from different fields are approached in a paper. The results by Chen, Arsenault, and Larivière (2015) show that “specialties see their interdisciplinarity increase steadily with the increase of each percentile
rank class” (p. 1042). Similar results have been published by Larivière and Gingras (2010) and Haustein, Larivière, and Börner (2014). In addition, more key words increase the probability of a paper to be discovered by a potential citer.

The seventh FIC in Table 2 is the number of words in the title of a paper (number of title words). This number of words might be referred to as one of many other “superficial factors that have nothing to do with the content [or quality] of the article” (Wesel, Wyatt, & Haaf, 2014, p. 1608). However, the results by Wesel et al. (2014) show that factors such as the number of title words can have a significant influence on citations (in certain fields). The results have been confirmed by other studies: the analysis by Letchford, Moat, and Preis (2015) “provides evidence that journals which publish papers with shorter titles receive more citations per paper. These results are consistent with the intriguing hypothesis that papers with shorter titles may be easier to understand, and hence attract more citations”. A negative correlation was also found by Guo, Ma, Shi, and Zong (2018) in economics. However, the negative correlation was visible only for the publication years until 2000: “the results show that correlation between title length and the number of citations is negative between 1956 and 2000, but becomes positive after 2000, when online searches became the predominant method for literature retrieval” (p. 1531).

The eighth FIC in Table 2 refers to two variables: the number of cited references and the number of linked cited references. Both variables are highly correlated but can deviate from one another especially in fields with a low coverage of the literature in WoS (or any other database such as Scopus). Whereas the number of cited references counts all cited references in a paper, linked cited references refer to only that part of cited references, which can be linked to a paper covered in the WoS database (i.e., a source item). Several studies revealed that the citation impact of papers is related to the number of references cited (see, e.g., Hegarty & Walton, 2012; Kostoff, 2007; Onodera & Yoshikane, 2014). Fok and Franses (2007) state that “longer articles with more references and also articles with more authors tend
to get more citations” (p. 386). Webster, Jonason, and Schember (2009) found that “log citations and log references were positively related … In other words, reference counts explained 19% of the variance in citation counts” (p. 356). Webster et al. (2009) mention two reasons for the correlation of number of cited references and citation counts: “First, review articles (e.g., theoretical reviews, meta-analyses) tend to have more citations than and are cited more frequently than typical empirical articles. Second, scientists are humans, and humans crave recognition for their work and often participate in reciprocal altruism” (p. 349).

As the results of Ahlgren, Colliander, and Sjögärde (2018) show not only the number of cited references, but also the share of cited references to papers within WoS is related to citation impact which justifies the consideration of both in the present study: the number of references and the number of linked references. Thus, the coverage of databases influences the times papers get cited (Ioannidis et al., 2016): if the coverage is low in a specific field, few links can be established between cited references and citing papers leading to low citation counts. Thus, one might imagine that if the complete literature from all fields would be included in a database and all cited references could be linked to a source item, field-specific differences in citation rates might be reduced.

The ninth FIC in Table 2 is the JIF which has been identified as one of the most important FIC in previous studies (see the FICs overview in Onodera & Yoshikane, 2014). The importance can be explained by the fact that “the more highly regarded a journal, the more likely it is that researchers will want to make use of its contents” (Meadows, 1998, p. 165). From the many studies investigating the JIF in the past, only a selected number can be presented here. Stanbrook and Redelmeier (2005) conclude that “journals act as passive conduits for scientific content, the impact of an article was strongly influenced by which journal published it. The journal impact factor appears to represent an accurate estimate of the relative ability of a journal to enhance article impact beyond the baseline contribution from article authors” (see also Knothe, 2006).
The results by Leimu and Koricheva (2005), Judge et al. (2007), Haslam and Koval (2010), and Ketzler and Zimmermann (2013) demonstrate the important role of the JIF as a FIC for management and ecological papers as well as papers in psychology and economics. Lariviere and Gingras (2010) conducted a quasi-experimental design to investigate the influence of the JIF on citations: they compared the citation impact of papers published twice in high and low impact journals (i.e., identical paper-pairs). The results show “that the journal in which papers are published have a strong influence on their citation rates, as duplicate papers published in high-impact journals obtain, on average, twice as many citations as their identical counterparts published in journals with lower impact factors” (p. 424). The results were confirmed by a study with a similar design conducted by Perneger (2010).

Lozano, Lariviere, and Gingras (2012) investigated the correlation between JIF and citations over time and revealed a time-dependent effect: “Throughout most of the 20th century, papers’ citation rates were increasingly linked to their respective journals’ IFs. However, since 1990, the advent of the digital age, the relation between IFs and paper citations has been weakening. This began first in physics, a field that was quick to make the transition into the electronic domain. Furthermore, since 1990 the overall proportion of highly cited papers coming from highly cited journals has been decreasing and, of these highly cited papers, the proportion not coming from highly cited journals has been increasing” (p. 2140). One might expect an increasing effect of the JIF on citations depending on the JIF values. That this is in fact the case was demonstrated by Milojević, Radicchi, and Bar-Ilan (2016): “The benefit of publishing in a journal with a higher IF value, instead of another with a lower one grows slowly as a function of the ratio of their IF values. For example, receiving more citations is granted in 90% of the cases only if the IF ratio is greater than 6, while targeting a journal with an IF 2 higher than another brings in marginal citation benefit”.
3.3 Statistical analyses

The need for field-normalization of bibliometric indicators has become a standard in bibliometrics, as it has been shown that papers are cited at different average levels in various fields (as outlined above). Therefore, the idea of considering fields as a cause of citation impact and perceiving them as treatments that have a causal influence on citation impact is very reasonable. However, looking at observed mean differences between fields gives little indication of the effect one field has on the citation impact compared to another, since many other FICs may play a role that confound the effect of the field (see section 3.2).

A causal inference is actually only possible in a classical randomized control group experiment, in which the units can be interchangeably assigned to the individual treatment groups. If the units are the papers and the treatments are the fields, the same paper cannot necessarily be assigned to all fields, which should be a prerequisite for a treatment. However, since interdisciplinary and transdisciplinary research strongly blurs the boundaries of fields, papers have cross-field characteristics especially when the focus is not on individual journals but on fields defined by a set of different journal-based categories (e.g., WoS subject categories). Thus, interchangeability of units is largely to be presupposed, which questions the use of fields (or journal sets) for normalizing citation impact in bibliometrics.

However, it is not yet possible to carry out a randomized group experiment (see AlShebli, Rahwan, & Woon, 2018): papers cannot be randomly assigned to fields for investigating the effect of the field on citation impact. But statistical concepts for causal inference, especially the so-called Rubin Causal Model (RCM, Rubin & Thomas, 1996), offer possibilities to draw causal conclusions for non-experimental designs as well. These manifold concepts are summarized under the term “propensity score matching” (Austin, Grootendorst, & Anderson, 2007; Rosenbaum & Rubin, 1983; Stuart, 2010). The method has already been used in bibliometrics (e.g., Colugnati, Firpo, de Castro, Sepulveda, & Salles, 2014; Farys & Wolbring, 2017; Mutz & Daniel, 2012b, 2012c; Mutz et al., 2017).
The “propensity score approach” is founded in Rubin’s potential outcome concept (Mutz et al., 2017, p. 2142). For an experiment with one treatment and one control for each unit i (here: a single paper), an individual treatment effect (here: a field) \( \alpha_i \) can be defined as the difference between what was observed if the unit i (i.e., a single paper) is exposed to treatment of the respective field \( Y_i(1) \), and what was observed if the same unit is exposed to control or all other fields \( Y_i(0) \): \( \alpha_i = Y_i(1) - Y_i(0) \), where \( Y_i(1) \) and \( Y_i(0) \) are the two potential outcomes for unit i. The fundamental problem of causal inference is that the individual causal effect \( \alpha_i \) is in principle not possible to estimate, as in an observational study only one potential outcome of a unit is realized through the unit treatment assignment \( T(0\ or\ 1) \) (Holland, 1986).

In a classical experiment, this problem is tackled by randomly assigning the units to the groups of interest. Individual effects for each single unit cannot be estimated in this setting, because only one potential outcome is realized in an experiment. However, at least the average effect between treatment and control can be thought of as an expected value, \( E \), (i.e., the mean of the individual effects between treatment and control across all units). This value can be interpreted causally, as average causal effect (ACE):

\[
ACE_i = E(Y_i(0)) - E(Y_i(1)) = E(Y_i(0|T=0)) - E(Y_i(1|T=1))
\]

(1)

With non-experimental designs, however, selection effects are to be expected, i.e. the assignment to the groups is not at random. For example, papers with many co-authors might be much more present in one field than in other fields, which might distort the average field-specific citation impact. Studies revealed that citation impact is associated with the number of co-authors and other FICs (see section 3.2). From this consideration, the general goal of this study is to model the selection or assignment process itself, i.e., to predict the probability of belonging to the treatment (here: fields) by means of a multiple logistic regression on an as
large as possible number of covariates. The assignment probability is nothing else than the propensi

24

s score for a unit (here: a single paper) to belong to a certain treatment. Units from treatment and control with the same propensity score have comparable values in their covariates (C). This makes it possible to match treatment and control groups in order to arrive at groups with a balanced distribution of the covariates. Eventually, this approach will result in a randomized control group experiment. Three assumptions are of central importance in this respect (Mutz et al., 2017, p. 2143):

1. **Strong ignorability** (Y(t) ┴ T|C): The potential outcomes Y(t) are independent of the field assignment as treatment T[t] given the set of covariates C. If all central covariates are considered, the selection process should be explained. Treatment effects can be interpreted causally (Imai & van Dyk, 2004). This assumption, however, cannot be verified in principle because in real experiments only one potential outcome of a unit occurs, either in treatment or control.

2. 0<p_i<1: The propensity scores p_i as probabilities should be greater than 0.0 and less than 1.0. This assumption would be violated, if the inclusion of a covariate leads to a complete separation of the groups. For example, if all papers in a certain field had a JIF of 0 and in the other fields JIF values were above 0, the propensity for that field would be 1.0 and for the other fields 0. This assumption can be checked by plotting the estimated propensity scores.

3. **Stable unit treatment value assumption** (SUTVA): It should be excluded in a study that the potential outcome of a paper published in a field is influenced by another paper being assigned to a certain field or not. Such possible interfering effects can be neglected in our case.

In the present study, there are more than two treatment groups, so that propensity score approaches for more than two groups have to be applied (Imai & van Dyk, 2004; Lopez & Gutman, 2017; McCaffrey et al., 2013; Mutz & Daniel, 2012b). The basis of the estimation of
propensity scores for multiple treatments is a multinomial multiple regression as an extension of a logistic regression to more than two groups. Instead of the usual group matching with regard to the propensity score, the “inverse-probability of treatment weighting” (IPW, McCaffrey et al., 2013, p. 3393) approach is used in this study. In this approach, the citation impact of a unit (here: paper) is weighted with the inverse of the corresponding propensity score of the unit in a field. While the mean value per field results in an unconfounded mean value of that field, the differences of the weighted mean values result in an average causal effect between the two fields (McCaffrey et al., 2013, p. 3393):

\[
\mu_t = \frac{\sum_{i=1}^{N} T_i[t]Y_i\omega_i[t]}{\sum_{i=1}^{N} T_i[t]\omega_i[t]}
\]

with weights \(\omega_i[t] = \frac{1}{p_T(C_i)}\),

where \(\mu_t\) is the unconfounded average of treatment or field \(t\) and \(T_i[t]\) is the treatment or field assignment variable. The variable is 1, if the respective paper belongs to field \(t\) and 0 if not. The average causal effect (ACE) of field 1 to field 2 is the estimated mean difference \(\mu_1 - \mu_2\).

This approach results in the following requirements for the selection of the covariates (see section 3.2):

- The covariates should influence both the treatment assignment and the citation impact.
- The covariates themselves should not be influenced by the treatment. This requirement may not be fulfilled in our study; that’s why we discuss it as a limitation in the discussion section.
- The covariates should not lead to complete separation of the treatments or fields.
- After the weighting with the propensity scores the differences between the fields in the covariates should vanish.
- A large set of covariates should be used in order to guarantee comparability of the fields.

The selected approach and the corresponding requirements results in the following procedure (Spreeuwenberg et al., 2010, p. 167):

- Effect estimation before correction.

- Balance check before correction: it is checked, whether the fields – as treatments – differ in the set of covariates.

- Estimation of propensity scores: the propensity scores are estimated by a multinomial regression of citation impact on a set of covariates.

- Check for overlap: it is checked, whether the distributions of scientific impact between disciplines overlap (diagnostic I).

- Balance check after correction (diagnostic II).

- Estimation of the causal effects after correction.

The statistical analyses were carried out with SAS 9.4 (SAS Institute Inc., 2012).

4 Results

The results section starts with reporting the diagnostic findings (in sections 4.1 and 4.2). Then, the estimated effects of the causal analyses – as the main results of the current study – are presented (in section 4.3).

4.1 Diagnostic I: assessing overlap of distributions

For a consistent estimation of the not confounded field-specific effects, there should be an overlap between the fields regarding the propensity scores. In principle, each paper should be able to occur in all fields, a complete separation of the groups should not be possible (see McCaffrey et al., 2013, p. 3406).
Since the propensity scores, i.e. the probabilities of paper allocations to all ten fields, are available for each paper of a particular field in this study, ten distributions can be represented for each field. Figure 1 shows the distribution of propensity scores for the ten WoS categories considered in this study as box plots. While the length of the boxes represents the interquartile distance (25% quartile to 75% quartile), the middle line in the boxes represents the median.
Figure 1. Boxplots for overlaps of multiple propensity scores (predicted probability) between the ten WoS categories included in this study (1=Biodiversity conservation, 2=Computer science, Artificial intelligence, 3=Communication, 4=Engineering, petroleum, 5=Family studies, 6=Geriatrics & gerontology, 7=Immunology, 8=Physics, particles & fields, 9=Rehabilitation, and 10=Spectroscopy)

It becomes clear in Figure 1 that the propensity scores – on average (median) – are highest for the fields to which the papers were assigned. For example, the second WoS category (Computer Science, Artificial Intelligence) has the highest average propensity score for the WoS category 2 (top picture on the right side of the figure). However, it also becomes clear for WoS category 2 that there is a greater than zero probability of potentially being able to be assigned to other WoS categories as well. Similar results as for WoS category 2 are observable for all WoS categories in Figure 1. Thus, the above-mentioned assumption of overlapping distributions is fulfilled for the dataset at hand.
We checked whether our results changed when fields on a higher aggregation level than the WoS subject categories are considered in the statistical analyses. The OECD assigns all WoS subject categories to main disciplinary categories. The results based on these categories are presented in Figure 2. It becomes clear that the interpretation which we delivered based on the WoS subject categories holds true for OECD subject categories. Although the propensity scores are on a high average level for the category the paper belongs to, there is also considerable overlap of distributions.

Figure 2. Boxplots for overlaps of multiple propensity scores between four OECD subject categories (papers published in the fourth category “Agricultural sciences” are not considered in this study)
4.2 Diagnostic II: balance check

The next diagnostic step was taken to examine whether the differences between the fields vanish when the propensity scores are considered as covariates in a regression. For the analysis of the count variables (e.g., number of pages) a negative binomial regression was calculated; for the analysis of the binary variables (e.g., a paper published by a European author or not) a logistic regression. The fields were included as additional covariates (as independent variables) in the model (ten WoS subject categories and four OECD subject categories, respectively) (Austin, 2009; Spreeuwenberg et al., 2010, p. 169). In other words, each model included a single FIC (e.g., number of pages) as dependent variable and the propensity scores and subject categories as independent variables.
Table 3. Propensity score check of covariates for WoS subject categories

| Variable                        | Propensity score? | F-Test WoS (F(9,294E3)) | Mean values – WoS subject categories |
|---------------------------------|-------------------|--------------------------|--------------------------------------|
|                                 |                   |                          | Biodiversity | Conservation | Communication | Computer Science, AI | Engineering, Petroleum | Family Studies | Geriatrics & Gerontology | Immunology | Physics, Particles & Fields | Rehabilitation | Spectroscopy |
| Number of pages                 | No                | 7,036.3                  | 11.7         | 12.6         | 18.2          | 8.0            | 14.8                  | 8.0            | 7.3                   | 11.5        | 9.5                  | 8.2          |               |
|                                 | Yes               | 124.9                    | 8.8          | 9.6          | 8.9           | 9.2            | 9.2                    | 8.8            | 8.8                   | 8.7         | 8.8                  | 8.8          |               |
| Number of author addresses      | No                | 1,958.8                  | 3.0          | 2.6          | 2.5           | 2.4            | 2.7                    | 3.7            | 3.7                   | 4.0         | 3.2                  | 3.1          |               |
|                                 | Yes               | 17.0                     | 3.1          | 3.3          | 3.3           | 3.3            | 3.2                    | 3.2            | 3.2                   | 3.2         | 3.2                  | 3.2          |               |
| Number of co-authors           | No                | 7,327.5                  | 3.0          | 2.7          | 2.0           | 2.7            | 2.6                    | 4.6            | 6.0                   | 9.2         | 3.4                  | 5.1          |               |
|                                 | Yes               | 317.7                    | 3.9          | 4.4          | 4.1           | 4.3            | 4.0                    | 4.1            | 4.5                   | 3.5         | 4.0                  | 3.9          |               |
| Number of cited references     | No                | 4,042.0                  | 38.6         | 20.3         | 39.5          | 12.6           | 37.2                   | 34.4           | 32.7                  | 28.9        | 31.1                 | 22.1         |               |
|                                 | Yes               | 69.2                     | 25.4         | 27.1         | 26.0          | 24.2           | 25.8                   | 25.8           | 25.5                  | 24.9        | 25.8                 | 24.9         |               |
| Number of linked cited references| No               | 1,1076.7                 | 18.9         | 8.2          | 12.3          | 3.3            | 16.3                   | 25.7           | 28.5                  | 19.7        | 16.9                 | 14.9         |               |
|                                 | Yes               | 66.2                     | 14.8         | 16.0         | 13.5          | 14.3           | 14.8                   | 15.5           | 15.4                  | 15.4        | 15.1                 | 15.6         |               |
| USA                             | No                | 1,838.9                  | .52          | .78          | .34           | .62            | .24                    | .52            | .59                   | .70         | .46                  | .74          |               |
|                                 | Yes               | 25.4                     | .64          | .56          | .58           | .61            | .62                    | .59            | .54                   | .56         | .61                  | .55          |               |
| Europe                          | No                | 794.1                    | .69          | .58          | .79           | .80            | .88                    | .62            | .59                   | .51         | .73                  | .55          |               |
|                                 | Yes               | 19.9                     | .75          | .72          | .75           | .68            | .74                    | .74            | .76                   | .76         | .76                  | .76          |               |
| Number of joined countries      | No                | 712.3                    | 1.3          | 1.2          | 1.1           | 1.1            | 1.2                    | 1.3            | 1.6                   | 1.1         | 1.1                  | 1.3          |               |
|                                 | Yes               | 3.8                      | 1.3          | 1.3          | 1.3           | 1.3            | 1.3                    | 1.3            | 1.3                   | 1.3         | 1.3                  | 1.3          |               |
| Number of title words           | No                | 8,296.9                  | 13.6         | 9.3          | 11.5          | 10.5           | 12.0                   | 12.9           | 14.6                  | 9.3         | 12.5                 | 12.9         |               |
|                                 | Yes               | 107.5                    | 11.9         | 12.1         | 12.2          | 11.7           | 11.8                   | 11.3           | 11.2                  | 11.5        | 11.6                 | 11.5         |               |
| Number of key words             | No                | 11,651.3                 | 5.9          | 1.8          | 3.5           | 1.09           | 5.0                    | 7.3            | 8.0                   | 4.7         | 5.2                  | 4.8          |               |
|                                 | Yes               | 48.1                     | 4.1          | 4.3          | 3.7           | 4.2            | 3.9                    | 4.0            | 4.0                   | 4.1         | 4.0                  | 4.0          |               |
| Journal Impact Factor           | No                | 14,357.2                 | 1.5          | 0.9          | 0.7           | 0.2            | 0.9                    | 2.3            | 3.8                   | 3.0         | 0.9                  | 1.6          |               |
|                                 | Yes               | 28.6                     | 1.7          | 1.7          | 1.6           | 1.6            | 1.7                    | 1.7            | 1.6                   | 1.7         | 1.7                  | 1.7          |               |
Both for the WoS categories (see Table 3) and for the OECD categories (see Table 4), a drastic but not complete reduction of the mean differences was observed, if the five propensity score variables were included in the regression analyses (grey background). For example, the mean values for the JIF are between 0.2 (“Engineering, Petroleum”) and 3.8 (“Immunology”) in Table 3. If the propensity scores are considered in the regression analyses, the mean JIFs in the fields vary only between 1.6 or 1.7.

The results in both tables can be interpreted also from the direct comparison of the F-tests before and after the correction. For example, the F-test for the regression analysis including “number of pages” reduces from 7,036.3 to 124.9, i.e. by 98.2% (see Table 3), if the propensity score variables are included as covariates in the analysis. Note that we waived statistical significance testing for the interpretation of the results in view of the rather large sample sizes (see section 3.1). All effects would be statistically significant anyway.

The results of the F-tests and the comparisons of the mean values for the FICs reveal that the fields are balanced with respect to the set of covariates. However, they are not completely balanced. There remain small differences, which might be traced back to the huge number of papers. This huge number leads to an enormous amount of the papers’ heterogeneity in our dataset which cannot be balanced out by the covariates.

Table 4. Propensity score check of covariates for four OECD subject categories

| Variable                      | Propensity score? | F-Test OECD (F(3,294E3)) | Natural sciences | Engineering and technology | Medical and Health sciences | Social sciences |
|-------------------------------|-------------------|---------------------------|------------------|---------------------------|----------------------------|----------------|
| Number of pages               | No                | 14,193.1                  | 12.0             | 9.5                       | 7.7                        | 14.8           |
|                               | Yes               | 252.6                     | 9.4              | 9.0                       | 8.9                        | 10.1           |
| Number of author addresses    | No                | 1,479.7                   | 3.3              | 2.9                       | 3.6                        | 2.7            |
|                               | Yes               | 10.3                      | 3.3              | 3.3                       | 3.2                        | 3.3            |
|                         | No         | 1,969.6 | 5.8 | 4.4 | 5.3 | 2.6 |
|-------------------------|------------|---------|-----|-----|-----|-----|
| Number of co-authors    | Yes        | 118.1   | 4.9 | 4.8 | 5.3 | 4.8 |
| Number of cited references| No        | 4,147.1 | 26.1| 23.1| 32.6| 37.2|
|                         | Yes        | 394.7   | 28.1| 25.5| 26.3| 31.2|
| Number of linked cited references| No | 12,733.3| 14.6| 13.1| 26.6| 16.3|
|                         | Yes        | 238.0   | 17.3| 15.7| 16.4| 19.0|
| USA                     | No         | 3,156.7 | .72 | .68 | .56 | .24 |
|                         | Yes        | 201.8   | .48 | .54 | .57 | .64 |
| Europe                  | No         | 858.7   | .62 | .61 | .55 | .88 |
|                         | Yes        | 23.5    | .73 | .72 | .74 | .66 |
| Number of joined countries| No       | 485.1   | 1.4 | 1.3 | 1.3 | 1.1 |
|                         | Yes        | 9.6     | 1.3 | 1.3 | 1.3 | 1.3 |
| Number of title words   | No         | 18,928.5| 9.7 | 12.4| 14.1| 12.0|
|                         | Yes        | 69.6    | 11.7| 11.7| 11.4| 11.7|
| Number of key words     | No         | 15,503.7| 3.5 | 4.2 | 7.5 | 5.0 |
|                         | Yes        | 207.9   | 4.8 | 4.5 | 4.5 | 5.2 |
| Journal Impact Factor   | No         | 15,392.2| 1.9 | 1.3 | 3.3 | 0.9 |
|                         | Yes        | 41.5    | 1.8 | 1.7 | 1.7 | 2.0 |

4.3 Estimated effects with and without correction for WoS subject categories

In the last step of the statistical analysis, the causal effects were estimated by weighting the citation data according to the propensity scores (see Table 3). While the mean value differences between the WoS subject categories represent the confounded effects before the correction (before IPW), the mean value differences after the correction (after IPW) represent the causal or unconfounded effects of the WoS subject categories, i.e., the “pure” effects of the WoS subject categories on citations. The “pure” effect relates to the condition that the fields do not differ any more with respect to the FICs included in the analyses.

In the interpretation of the results in Table 3, not only the mean value differences between the subject categories after IPW can be considered, but also the changes of the mean values before and after the weighting for single subject categories. Taken as a whole, the mean differences between the WoS categories in Table 3 become smaller due to the correction. While the subject categories “Rehabilitation” and “Computer Science, AI” show no major changes in the mean values, the subject categories “Communication” and “Spectroscopy” have significantly higher citation impact values after correction. For the
subject category “Communication”, the citation impact value rises from 23.65 to 48.37; for the subject category “Spectroscopy” from 18.42 to 37.92. The increase in the citation impact values for these fields points out that the citation impact is undervalued, if the uncorrected or confounded citation data are used. Since the results in Table 5 for “Immunology” demonstrate a decrease in citation impact values by considering the weighting (from 40.78 to 34.07), the impact is overvalued in this case.

Table 5. Effects of WoS subject categories on total citations before and after IPW

| WoS subject category        | Before IPW | After IPW |
|----------------------------|------------|-----------|
|                            | Number of papers | Mean | Standard deviation | Mean | Standard deviation |
| Biodiversity Conservation   | 10,187     | 31.95 | 51.63 | 37.84 | 87.32 |
| Communication              | 48,173     | 23.65 | 162.94 | 48.37 | 150.13 |
| Computer Science, AI       | 6,003      | 24.53 | 40.42 | 23.18 | 36.27 |
| Engineering, Petroleum     | 6,016      | 5.89  | 14.14 | 17.80 | 25.81 |
| Family Studies             | 5,983      | 27.03 | 49.25 | 19.25 | 34.07 |
| Geriatrics & Gerontology   | 12,193     | 35.75 | 69.35 | 30.00 | 47.12 |
| Immunology                 | 98,621     | 40.78 | 69.76 | 34.07 | 65.24 |
| Physics, Particles & Fields| 53,218     | 23.31 | 53.42 | 28.19 | 170.70 |
| Rehabilitation             | 18,197     | 24.53 | 39.75 | 24.56 | 44.60 |
| Spectroscopy               | 35,508     | 18.42 | 36.62 | 37.92 | 141.40 |

Note. For IPW not only a weighted mean, but also a weighted standard deviation was calculated, where the divisor is the sum of weights.

4.4 Estimated effects with and without correction for OECD categories

There are also differences with regard to the OECD categories (see Table 6). While there are no differences for the OECD Category “Natural sciences” before and after IPW, the OECD Category “Engineering and technology” is clearly undervalued in the observed citation
data. The mean value rises from 17.61 (before IPW) to 36.76 (after IPW). The OECD Category “Medical and health sciences” is overvalued (38.01 before IPW, 29.63 after IPW).

Table 6. Effects of OECD categories on total citations before and after IPW

| OECD Category                  | Number of papers | Before IPW | After IPW |
|-------------------------------|------------------|------------|-----------|
|                               | Mean             | Standard deviation | Mean     | Standard deviation |
| Natural sciences              | 111,578          | 24.24      | 114.34    | 24.49      | 90.25       |
| Engineering and technology    | 47,527           | 17.61      | 35.46     | 36.76      | 141.63     |
| Medical and health sciences   | 129,011          | 38.01      | 66.55     | 29.63      | 58.64      |
| Social sciences               | 5,983            | 27.03      | 49.24     | 21.95      | 37.69      |

Note. For IPW not only a weighted mean, but also a weighted standard deviation was calculated, where the divisor is the sum of weights.

5 Discussion

Field normalization of citation indicators is one of the gold standards in bibliometrics (Hicks et al., 2015). It is widely accepted among bibliometricians that the interpretation of raw citation counts across the boundaries of different fields makes little sense given the different field-specific citation levels. According to Waltman (2016), there often is a need to compare papers from different fields and “normalized citation impact indicators have been developed to make such comparisons. The idea of these indicators is to correct as much as possible for the effect of variables that one does not want to influence the outcomes of a citation analysis, such as the field” (p. 375). For Waltman and van Eck (2019), “differences between fields lead to distortions in scientometric indicators. One could think of this in terms of signal and noise. Scientometric indicators provide a signal of concepts such as productivity or scientific impact, but they are also affected by noise. This noise may partly be due to differences between fields, for instance, differences in publication, collaboration, and citation
practices. Field normalization aims to remove this noise while maintaining the signal” (p. 282).

Despite the observed citation impact differences between fields, the question remains how strong fields (however defined) influence the citation impact of papers. The question arises because of two reasons: (1) studies which have investigated reasons to cite or citation functions revealed that many FICs exist – the field is only one FIC besides many others. However, as the field is confounded with the other FICs, additional citation effects beyond the field itself are covered by field-normalization. According to Ioannidis et al. (2016), “citations received by a single paper may depend on multiple factors beyond pure merit. These include the scientific field where the paper belongs (different fields have different numbers of publishing and citing scientists and citing cultures, and thus different citation density), the age (how long ago it was published), the type of document (article, review, letter, editorial), and the coverage of the database where citations are counted”. Given these and many other FICs, the probability is high that the FIC “field” is confounded by one or several other FICs (as demonstrated using the FIC “number of co-authors” as an example in the introduction section).

(2) It is not clear how a field can be properly defined (separated) in bibliometrics. Fields can be invisible colleges (see above), organizational units, and common research topics (among other things). In bibliometrics, field-normalized citation scores are calculated based on different field-classification schemes, which are based on journal sets (differently composed), citation relations, and human-based assignments. According to Waltman and van Eck (2013a), “normalization based on a field classification system has a number of limitations. First, the idea of science being subdivided into a number of clearly delineated fields is artificial. In reality, boundaries between fields tend to be rather fuzzy. Second, fields can be defined at different levels of granularity, and given a certain level at which one has defined one’s fields, it is always possible to go one level deeper and to define subfields at this
deeper level. It is quite well possible that the subfields within a single field differ significantly from each other in terms of citation practices … Hence, in many cases, it is not clear to what extent fields can be regarded as homogeneous entities” (p. 700).

Ioannidis et al. (2016) explain the problem of usual field-normalization in bibliometrics as follows: “a major challenge is how to define scientific fields for normalization. Fields have been categorized in the past on the basis of journals or library categories. Within-field citations are usually denser than between-field citations. However, no field is isolated, and between-field communication is increasingly common nowadays … In some areas, the boundaries between fields seem to become less distinct. Different categorizations may segregate science to anywhere between a dozen … to several thousands of disciplines or specialties … Fields may be defined a priori (e.g., based on the journals where research gets published) or dynamically (e.g., based on citing or cited papers or on the references of citing or cited papers) … The rationale is that citing papers consider cited papers relevant to their work, so they belong to the same field. Obviously, this is not true for all citations – e.g., some methods (statistical, laboratory, or other) can be used by multiple unrelated fields, and there are also substantive interdisciplinary references”.

Thus, if it is not possible to properly delimit the papers belonging to a single field in bibliometrics, how can we decide that the FIC “field” should be considered in measuring citation impact? A possible valid assignment of papers to fields might diminish field-specific differences in impact scores, which are observable to various extents based on the field-categorization schemes used by bibliometricians today. In contrast, the number of co-authors – another FIC besides the field – is clearly defined and realizable in publication data but is usually not considered in normalization approaches of citation impact. When we started this study, we were confronted with the problem of selecting a field-categorization scheme for this study. Journal sets are one of the most popular field-categorization schemes for field-normalization. Thus, we took this scheme as basis for investigating the causal effects of fields
on citations in the current study. In order to distill the “pure” effect of the field (defined by the scheme) on citation impact, this contribution is devoted to statistical concepts of causal inference.

Using the MPG in-house version of the WoS database, we selected ten WoS subject categories (fields) which have as large as possible differences in mean citation rates. We were interested in the question whether the differences can be traced back to other FICs (than fields) or are “pure” field effects. We considered number of pages, number of co-authors, number of author addresses, number of joined countries, the binary national variables “USA” and “Europe”, number of keywords, number of title words, number of (linked) cited references, and JIF as FICs in this study. Although many more FICs have been proposed and investigated in previous studies, we included only a reduced set covering frequently studied FICs and FICs with a proposed significant influence on citation impact. However, the selection of the FICs might be a limitation of the study which is explained in the following (more FICs could have been included).

Although many studies have been conducted previously on FICs and the relationship between citation impact and fields, we are not aware of any other studies applying the propensity score matching approach being a family including many different procedures. This approach can be used for causal inferences comparable to an experimental design, however, with less methodological stringency due to the non-experimental nature of the used data. Most of the studies on FICs and the relationship between citation impact and fields were based on correlations or ordinary regression analyses which cannot make any claim to causality. Propensity score matching and the corresponding concept of causal inference, called Rubin Causal Model, defines statistical conditions being fulfilled to make causal statements. In propensity score matching (i.e. inverse propensity weighting) groups (here: fields) are established, which do not differ anymore in central characteristics except for field. Very convincing for the advantage of propensity score matching are studies comparing the quasi-
experimental design with randomized control group experiments on the same topic. The results of these studies revealed that the treatment effects did not differ when a quasi-experimental design with propensity score matching was compared with a real experiment (e.g., Cook, Shadish, & Wong, 2008).

In one of two diagnostic steps in our statistical analyses, we considered the calculated propensity scores as covariates in regression analyses to examine whether the differences between the fields vanish. As the results showed, the differences did not completely vanish, but were significantly reduced. We received similar results when we calculated mean value differences of the fields after IPW representing the causal or unconfounded field effects on citations: the fields still differ with respect to the FICs included in the analyses.

Our results can be interpreted in different ways. One possible interpretation is that field-differences in citations exist which are relatively independent of other FICs. In this case, normalization of citation impact would be reasonable although our results would question the focus on fields in normalization approaches. Other FICs have (significant) effects on citation impact; a result which has been confirmed by other studies (e.g., Yu et al., 2014). However, our results could also be interpreted such that field-normalization does not only capture pure field effects but also other effects (FICs-related) which are desirable to include in normalization procedures. Another possible interpretation of our results is that field-differences could vanish in principle, but the following three potential limitations of the study prevented a clear-cut result:

- **Selection of covariates:** For the propensity score matching approach, an as large as possible number of covariates is required to establish comparability of the groups (here: fields). We included important covariates in this study, but did not consider many other covariates proposed in the past. For example, covariates explicitly referring to the quality of papers – independent of citations – were not available for our data at hand (e.g., quality assessments by peers). Since the normative theory of
citing introduced by Merton (1973) bases citation decisions exclusively on the quality concept, an important factor is not considered in this study.

- **Balance of the groups:** Although a strong reduction of the mean differences between the fields in the covariates could be observed, the differences do not vanish completely. Given the rather large sample size, this is not surprising: more papers mean a greater heterogeneity among the papers which might be more difficult to capture with the set of covariates considered in this study. The consideration of more covariates might be a possible solution for this problem (in future studies).

- **Impact of fields on covariates:** We included covariates in this study which have been frequently investigated in previous studies as FICs. It is unproblematic for applying the propensity score matching approach if the covariates are correlated both with citation counts and subject categories (confounding the causal relationship). But the covariates should not themselves be influenced by the subject categories. For example, the higher the number of cited references of papers, the higher the citation counts. If in field A papers with many cited references are more frequent than in field B, then the observed higher citations in field A might be also due to differences in the number of cited references. Thus, for the FIC “cited references” it might be plausible to argue that the field does not influence it. The same might be true for the other FICs included in this study – except one (the JIF, see above).

Since our study is based on a rather large sample size with great heterogeneity (see above), we can imagine, however, that this critical point might not distort our results.

- **Field-categorization scheme:** We used WoS subject categories as field-categorization scheme in this study. In section 2, we outlined that several schemes exist in bibliometrics (and no standard scheme). Since the schemes are conceptually very different, a considerable impact of the schemes on field-specific impact scores might be possible. Thus, it would be interesting to include not only one scheme – as we did –
but several schemes in future studies (e.g., based on citation relations or human-based
dfield-assignments). Then, the influence of the schemes – independent of the field –
could be estimated on the results.

Although the results of our study point out that field-effects on citations still exist even
if FICs are considered, the reduction of field-effects might question the use of field-
categorization schemes for normalizing citations. Since citation impact can be alternatively
measured in research evaluation, these alternatives might be preferred. Two alternatives have
been proposed in previous years which allow a kind of benchmarking “that is needed to
enhance comparability across diverse scientists, fields, papers, time periods, and so forth”
(Ioannidis et al., 2016):

(1) The first alternative are citing-side indicators (Zitt & Small, 2008) which focus on
the basic premise of normalization that “not all citations are equal” (Ioannidis et al., 2016). In
this study, we targeted the cited-side indicator’s approach, which started from the premise that
not all papers are equal (but are embedded in a certain field-specific publication and citation
culture). With citing-side indicators, every citation of a paper is normalized considering the
citation density in the citing paper (measured on the basis of its cited references or its
publishing journal). Thus, citing-side indicators are “normalised citation impact indicators
without defining fields in an explicit way” (Wilsdon et al., 2015, p. 32). In recent years,
several variants of citing-side indicators have been proposed. The results by Waltman and van
Eck (2013a) show that they “may yield more accurate results” (p. 699) than usual field-
normalized indicators, Bornmann and Marx (2015) partly confirmed the results. However,
Bornmann and Haunschild (2017) found high correlations between cited-side and citing-side
indicators and no significant difference between the correlation of cited-side indicators and
F1000Prime scores compared with citing-side indicators and F1000Prime scores for the bio-
medical literature.
(2) Another alternative to field-normalization in evaluative bibliometrics is to use non-normalized indicators and to “contextualize these indicators with additional information that enables evaluators to take into account the effect of field differences … For instance, to compare the productivity of researchers working in different fields, one could present non-normalized productivity indicators (e.g., total publication or citation counts during a certain time period) for each of the researchers to be compared. One could then contextualize these indicators by selecting for each researcher a number of relevant peers working in the same field and by also presenting the productivity indicators for these peers. In this way, each researcher’s productivity can be assessed in the context of the productivity of a number of colleagues who have a reasonably similar scientific profile” (Waltman & van Eck, 2019, p. 295). The use of non-normalized indicators would also agree to the demand to use simple indicators in evaluative bibliometrics: “there is often a demand for simple measures because they are easier to use and can facilitate comparisons” (University of Waterloo Working Group on Bibliometrics, 2016, p. 2). These indicators are better understandable by experts in the evaluated fields and facilitate proper interpretation of the bibliometric results.

We would like to emphasize that we did not produce conclusive results for finally deciding on the use of field-normalized indicators in evaluative bibliometrics. Our study might be a starting point not to understand field-normalization as unquestioned standard, but to investigate this issue in future studies.
Acknowledgements

The bibliometric data used in this paper are from an in-house database developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics (Philadelphia, Pennsylvania, USA).
References

Abramo, G., Cicero, T., & D’Angelo, C. A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics, 5*(4), 659-667. doi: 10.1016/j.joi.2011.06.004.

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. doi: 10.1002/asi.23986.

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications, 9*, 5163. doi: 10.1038/s41467-018-07634-8.

Andersen, H. (2016). Collaboration, interdisciplinarity, and the epistemology of contemporary science. *Studies in History and Philosophy of Science Part A, 56*, 1-10. doi: 10.1016/j.shpsa.2015.10.006.

Antonakis, J., Bastardoz, N., Liu, Y. H., & Schriesheim, C. A. (2014). What makes articles highly cited? *Leadership Quarterly, 25*(1), 152-179. doi: 10.1016/j.leaqua.2013.10.014.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107. doi: 10.1002/sim.3697.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine, 26*(4), 734-753. doi: 10.1002/sim.2580.

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

Benavent-Pérez, M., Gorraiz, J., Gumpenberger, C., & Moya-Anegón, F. (2012). The different flavors of research collaboration: a case study of their influence on university excellence in four world regions. *Scientometrics, 93*(1), 41-58. doi: 10.1007/s11192-012-0638-4.

Bornmann, L., & Daniel, H.-D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59*(11), 1841-1852. doi: 10.1002/asi.20901.

Bornmann, L., & Haunschild, R. (2017). Relative Citation Ratio (RCR): An Empirical Attempt to Study a New Field-Normalized Bibliometric Indicator. *Journal of the Association for Information Science and Technology, 68*(4), 1064-1067. doi: 10.1002/asi.23729.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics, 9*(2), 408-418.

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

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*, 93-102. doi: 10.3354/esep00084.
Bornmann, L., Stefaner, M., de Moya Anegon, F., & Mutz, R. (2014). What is the effect of country-specific characteristics on the research performance of scientific institutions? Using multi-level statistical models to rank and map universities and research-focused institutions worldwide. *Journal of Informetrics, 8*(3), 581-593. doi: 10.1016/j.joi.2014.04.008.

Bornmann, L., Wagner, C., & Leydesdorff, L. (2018). The geography of references in elite articles: What countries contribute to the archives of knowledge. *PLoS ONE, 13*(3), e0194805. doi: 10.1371/journal.pone.0194805.

Bornmann, L., & Wohlrabe, K. (2019). Normalisation of Citation Impact in Economics. *Scientometrics, 120*(2), 841-884. doi: 10.1007/s11192-019-03140-w.

Boyack, K. W. (2004). Mapping knowledge domains: characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5192-5199.

Chen, S., Arsenault, C., & Larivière, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics, 9*(4), 1034-1046. doi: 10.1016/j.joi.2015.09.003.

Colugnati, F. A. B., Firpo, S., de Castro, P. F. D., Sepulveda, J. E., & Salles, S. L. M. (2014). A propensity score approach in the impact evaluation on scientific production in Brazilian biodiversity research: the BIOTA Program. *Scientometrics, 101*(1), 85-107. doi: 10.1007/s11192-014-1397-1.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750. doi: 10.1002/pam.20375.

Crespo, J. A., Li, Y. R., & Ruiz-Castillo, J. (2013). The Measurement of the Effect on Citation Inequality of Differences in Citation Practices across Scientific Fields. *Plos One, 8*(3). doi: 10.1371/journal.pone.0058727.

Di Vaio, G., Waldenström, D., & Weisdorf, J. (2012). Citation success: evidence from economic history journal publications. *Explorations in Economic History, 49*(1), 92-104. doi: 10.1016/eeh.2011.10.002.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64*(5), 1055-1064. doi: 10.1002/asi.22806.

Didegah, F., & Thelwall, M. (2014). Article Properties Associating with the Citation Impact of Individual Articles in the Social Sciences. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden “Context Counts: Pathways to Master Big and Little Data”* (pp. 169-175). Leiden, the Netherlands: University of Leiden.

Dong, Y., Ma, H., Tang, J., & Wang, K. (2018). Collaboration Diversity and Scientific Impact. Retrieved September 5, 2018, from [https://arxiv.org/abs/1806.03694](https://arxiv.org/abs/1806.03694).

Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. *Plos One, 8*(2). doi: 10.1371/journal.pone.0049476.

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. doi: 10.1002/asi.23802.

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

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113-122.

Gillmor, C. S. (1975). Citation characteristics of JATP literature. *Journal of Atmospheric and Terrestrial Physics, 37*(11), 1401-1404.
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. doi: 10.1007/s11192-017-2593-6.

Glänzel, W. (2013). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlsberger & H. F. Moed (Eds.), *Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference* (Vol. 1, pp. 109-122). Vienna, Austria.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics, 67*(2), 263-277.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14*(2), 123-127.

Guo, F., Ma, C., Shi, Q., & Zong, Q. (2018). Succinct effect or informative effect: the relationship between title length and the number of citations. *Scientometrics, 116*(3), 1531-1539. doi: 10.1007/s11192-018-2805-8.

Haddow, G., & Noyons, E. (2013). Misfits? research classification in research evaluation: Visualizing journal content within fields of research codes. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger & H. F. Moed (Eds.), *Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference* (Vol. 2, pp. 1210-1224). Vienna, Austria.

Hammarfelt, B. (2018). What is a discipline? The conceptualization of research areas and their operationalization in bibliometric research. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2018 Leiden “Science, Technology and Innovation indicators in transition”*. Leiden, the Netherlands: University of Leiden.

Hanel, P. H. P., & Haase, J. (2017). Predictors of Citation Rate in Psychology: Inconclusive Influence of Effect and Sample Size. *Frontiers in Psychology, 8*. doi: 10.3389/fpsyg.2017.01160.

Haslam, N., & Koval, P. (2010). Predicting long-term citation impact of articles in social and personality psychology. *Psychological Reports, 106*(3), 891-900. doi: 10.2466/pr0.106.3.891-900.

Haunschild, R., Marx, W., French, B., & Bornmann, L. (2018). Relationship between field-normalized indicators calculated with different approaches of field-categorization. Paper presented at the International Conference on Science and Technology Indicators (STI 2018), Leiden, The Netherlands. https://openaccess.leidenuniv.nl/handle/1887/65267

Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. *Journal of Informetrics, 12*(2), 436-447. doi: 10.1016/j.joi.2018.03.004.

Haustein, S., Larivière, V., & Börner, K. (2014). Long-distance interdisciplinary research leads to higher citation impact. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden “Context Counts: Pathways to Master Big and Little Data”* (pp. 256-259). Leiden, the Netherlands: University of Leiden.

Hegarty, P., & Walton, Z. (2012). The consequences of predicting scientific impact in psychology using Journal Impact Factors. *Perspectives on Psychological Science, 7*(1), 72-78. doi: 10.1177/1745691611429356.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(467), 854-866. doi: 10.1198/016214504000001187.

Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation Metrics: A Primer on How (Not) to Normalize. *PLoS Biol, 14*(9), e1002542. doi: 10.1371/journal.pbio.1002542.

Iribarren-Maestro, I., Lascurain-Sanchez, M. L., & Sanz-Casado, E. (2007). Are multi-authorship and visibility related? Study of ten research areas at Carlos III University of Madrid. In D. Torres-Salinas & H. F. Moed (Eds.), *Proceedings of the 11th Conference of the International Society for Scientometrics and Informetrics* (Vol. 1, pp. 401-407). Madrid, Spain: Spanish Research Council (CSIC).

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *Science, 322*(5905), 1259-1262. doi: 10.1126/science.1158357.

Judge, T., Cable, D., Colbert, A., & Rynes, S. (2007). What causes a management article to be cited - article, author, or journal? *The Academy of Management Journal (AMJ), 50*(3), 491-506.

Ketzler, R., & Zimmermann, K. (2013). A citation-analysis of economic research institutes. *Scientometrics, 95*(3), 1095-1112. doi: 10.1007/s11192-012-0850-2.

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998. doi: 10.1002/asi.23734.

Knothe, G. (2006). Comparative citation analysis of duplicate or highly related publications. *Journal of the American Society for Information Science and Technology, 57*(13), 1830-1839.

Kostoff, R. (2007). The difference between highly and poorly cited medical articles in the journal Lancet. *Scientometrics, 72*(3), 513-520.

Kronman, U., Gunnarsson, M., & Karlsson, S. (2010). *The bibliometric database at the Swedish Research Council — contents, methods and indicators*. Stockholm, Sweden: Swedish Research Council.

Larivière, V., & Gingras, Y. (2010). The impact factor’s Matthew effect: a natural experiment in bibliometrics. *Journal of the Association for Information Science and Technology, 61*(2), 424-427.

Larivière, V., & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Larivière, V., Gong, K., & Sugimoto, C. R. (2018). Citations strength begins at home. *Nature, 564*(7735), S70-S71. doi: 10.1038/d41586-018-07695-1.

Lawani, S. M. (1986). Some bibliometric correlates of quality in scientific research. *Scientometrics, 9*(1-2), 13-25.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20*(1), 28-32.

Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. *Royal Society Open Science, 2*(8). doi: 10.1098/rsos.150266.

Leydesdorff, L., & Milojević, S. (2015). The Citation Impact of German Sociology Journals: Some Problems with the Use of Scientometric Indicators in Journal and Research Evaluations. *Soziale Welt, 66*(2), 193-204.

Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—a
state-of-the-art report. *Scientometrics*, 109(3), 2129–2150. doi: 10.1007/s11192-016-2150-8.

Liu, M. X. (1993). A study of citing motivation of Chinese scientists. *Journal of Information Science*, 19(1), 13-23.

Lopez, M. J., & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, 32(3), 432-454. doi: 10.1214/17-sts612.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, 63(11), 2140-2145. doi: 10.1002/asi.22731.

Lundberg, J. (2007). Lifting the crown - citation z-score. *Journal of Informetrics*, 1(2), 145-154.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: a study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1), 1-12.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19), 3388-3414. doi: 10.1002/sim.5753.

McCain, K. W. (2014). Obliteration by Incorporation. In C. Blaise & R. S. Cassidy (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 129-149). Cambridge, MA, USA: MIT Press.

Meadows, A. J. (1998). *Communicating research*. London, UK: Academic Press.

Merton, R. K. (1973). *The sociology of science: theoretical and empirical investigations*. Chicago, IL, USA: University of Chicago Press.

Merton, R. K. (1988). The Matthew effect in science, II: cumulative advantage and the symbolism of intellectual property. *ISIS*, 79(4), 606-623.

Milojević, S., Radicchi, F., & Bar-Ilan, J. (2016). Validity of Basing Authors' Publication Strategy on Journal Impact Factors. Retrieved September, 1, 2016, from http://arxiv.org/abs/1607.03179

Mingers, J., & Leydesdorff, L. (2015). A Review of Theory and Practice in Scientometrics. *European Journal of Operational Research*, 246(1), 1-19.

Moed, H. F. (2017). *Applied Evaluative Informetrics*. Heidelberg, Germany: Springer.

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

Mutz, R., & Daniel, H.-D. (2019). How to consider fractional counting and field normalization in the statistical modeling of bibliometric data: A multilevel Poisson regression approach. *Journal of Informetrics*, 13(2), 643-657. doi: 10.1016/j.joi.2019.03.007.

Mutz, R., & Daniel, H. D. (2012b). The generalized propensity score methodology for estimating unbiased journal impact factors. *Scientometrics*, 92(2), 377-390. doi: 10.1007/s11192-012-0670-4.

Mutz, R., & Daniel, H. D. (2012c). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics*, 6(2), 169-176. doi: 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. doi: 10.1002/asi.23701.
Onodera, N., & Yoshikane, F. (2014). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology, 66*(4), 739–764. doi: 10.1002/asi.23209.

Pasterkamp, G., Rotmans, J. I., de Kleijn, D. V. P., & Borst, C. (2007). Citation frequency: a biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics, 70*(1), 153-165.

Perneger, T. V. (2010). Citation analysis of identical consensus statements revealed journal-related bias. *Journal of Clinical Epidemiology, 63*(6), 660-664. doi: 10.1016/j.jclinepi.2009.09.012.

Purkayasthaa, A., Palmaroa, E., Falk-Krzesinskib, H. J., & Baas, J. (2019). Comparison of two article-level, feld-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). *Journal of Informetrics, 13*(2), 635-642.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E, 83*(4). doi: 10.1103/PhysRevE.83.046116.

Riviera, E. (2013). Scientific communities as autopoietic systems: The reproductive function of citations. *Journal of the American Society for Information Science and Technology, 64*(7), 1442-1453. doi: 10.1002/asi.22826.

Robson, B. J., & Mousquès, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling & Software, 75*, 94-104. doi: 10.1016/j.envsoft.2015.10.007.

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. doi: 10.1093/biomet/70.1.41.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*(1), 249-264. doi: 10.2307/2533160.

Sandström, U. (2014). *Bibliometric evaluation of SEPA-funded large research programs 2003–2013*. Stockholm, Sweden: Swedish Environmental Protection Agency.

SAS Institute Inc. (2012). *SAS/STAT® 12.1 User’s Guide*. Cary, NC: SAS Institute Inc.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hагенаарс, J. A., Busschbach, J. J. V., Andrea, H., . . . Stijnen, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms An introduction from a case study in Mental Health. *Medical Care, 48*(2), 166-174. doi: 10.1097/MLR.0b013e3181c1328f.

Stanbrook, M. B., & Redelmeier, D. A. (2005, September 17, 2005). *Text vs context: the influence of the journal on article impact*. Paper presented at the 5th International Congress on Peer Review and Biomedical Publication, Chicago, IL, USA.

Stanek, K. Z. (2008). How long should an astronomical paper be to increase its Impact? Retrieved September 22, 2008, from http://arxiv.org/abs/0809.0692

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics, 4*(2), 194-200. doi: 10.1016/j.joi.2009.12.001.

Stuart, E. A. (2010). Matching methods for causal Inference: A review and a look forward. *Statistical Science, 25*(1), 1-21. doi: 10.1214/09-STS313.

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation, 71*(4), 775-794. doi: 10.1108/Jd-06-2014-0082.

Taborsky, M. (2009). Biased citation practice and taxonomic parochialism. *Ethology, 115*(2), 105-111. doi: 10.1111/j.1439-0310.2009.01610.x.

Tabamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*(1), 203-216. doi: 10.1016/j.joi.2018.01.002.
Tahamtan, I., & Bornmann, L. (in press). What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018. *Scientometrics*. doi: 10.1007/s11192-019-03243-4.

Tahamtan, I., Safipour Afshar, A., & Ahamdizadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics, 107*(3), 1195-1225. doi: 10.1007/s11192-016-1889-2.

Teplitskiy, M., & Bakanic, V. (2016). Do Peer Reviews Predict Impact? Evidence from the American Sociological Review, 1978 to 1982. *Socius, 2*, 2378023116640278. doi: 10.1177/2378023116640278.

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

University of Waterloo Working Group on Bibliometrics. (2016). *White Paper on Bibliometrics, Measuring Research Outputs through Bibliometrics*. Waterloo, Ontario.

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

van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *Plos One, 8*(4). doi: 10.1371/journal.pone.0062395.

Walstad, A. (2002). Science as a market process. *The Independent Review, 7*(1), 5-45.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419-2432.

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. doi: 10.1002/asi.22748.

Waltman, L., & van Eck, N. J. (2013a). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics, 96*(3), 699-716. doi: 10.1007/s11192-012-0913-4.

Waltman, L., & van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics, 7*(4), 833-849.

Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Handbook of Science and Technology Indicators* (pp. 281-300). Heidelberg, Germany: Springer.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics, 5*(1), 37-47. doi: 10.1016/j.joi.2010.08.001.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science, 342*(6154), 127-132. doi: 10.1126/science.1237825.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872. doi: 10.1007/s11192-012-0775-9.

Wang, Q., & Waltman, L. (2016). Large-Scale Analysis of the Accuracy of the Journal Classification Systems of Web of Science and Scopus. *Journal of Informetrics, 10*(2), 347-364.

Webster, G. D., Jonason, P. K., & Schember, T. O. (2009). Hot topics and popular papers in evolutionary psychology: analyses of title words and citation counts in *Evolution and Human Behavior, 1979-2008*. *Evolutionary Psychology, 7*(3), 348-362.
Wesel, M., Wyatt, S., & Haaf, J. (2014). What a difference a colon makes: how superficial factors influence subsequent citation. *Scientometrics, 98*(3), 1601–1615. doi: 10.1007/s11192-013-1154-x.

Whitfield, J. (2008). Group theory. *Nature, 455*, 720-723.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: Higher Education Funding Council for England (HEFCE).

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)*. London, UK: Higher Education Funding Council for England (HEFCE).

Wray, K. B. (2006). Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science Part A, 37*(3), 505-514.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*(2), 1233-1252. doi: 10.1007/s11192-014-1279-6.

Ziman, J. (1996). "Postacademic Science": Constructing Knowledge with Networks and Norms. *Science Studies, 9*(1), 67-80.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856-1860. doi: 10.1002/asi.20880.