AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Andersen, J. P., & Nielsen, M. W. (2018). Google Scholar and Web of Science: Examining gender differences in citation coverage across five scientific disciplines. *Journal of Informetrics*, 12(3), 950-959. <https://doi.org/10.1016/j.joi.2018.07.010>

Publication metadata

Title:	Google Scholar and Web of Science: Examining gender differences in citation coverage across five scientific disciplines
Author(s):	Jens Peter Andersen & Mathias Wullum Nielsen
Journal:	<i>Journal of Informetrics</i> , 12(3), 950-959
DOI/Link:	https://doi.org/10.1016/j.joi.2018.07.010
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Google Scholar and Web of Science: Examining gender differences in citation coverage across five scientific disciplines^{*}

Jens Peter Andersen^{a,*}, Mathias Wullum Nielsen^a

*^aDanish Centre for Studies on Research and Research Policy. Aarhus University.
Bartholins Alle 7, Dk-8000 Aarhus C. Denmark.*

Abstract

Many studies demonstrate differences in the coverage of citing publications in Google Scholar (GS) and Web of Science (WoS). Here, we examine to what extent citation data from the two databases reflect the scholarly impact of women and men differently. Our conjecture is that WoS carries an indirect gender bias in its selection criteria for citation sources that GS avoids due to criteria that are more inclusive. Using a sample of 1,250 U.S. researchers in Sociology, Political Science, Economics, Cardiology and Chemistry, we examine gender differences in the average citation coverage of the two databases. We also calculate database-specific h-indices for all authors in the sample. In repeated simulations of hiring scenarios, we use these indices to examine whether women's appointment rates increase if hiring decisions rely on data from GS in lieu of WoS. We find no systematic gender differences in the citation coverage of the two databases. Further, our results indicate marginal to non-existing effects of database selection on women's success-rates in the simulations. In line with the existing literature, we find the citation coverage in WoS to be largest in Cardiology and Chemistry and smallest in Political Science and Sociology.

^{*}The research presented here is the continuation of a research-in-progress paper presented at the 16th International Conference of the International Society for Scientometrics and Informetrics in Wuhan, China (Andersen & Nielsen, 2017). It has been substantially appended and modified since then.

^{*}Corresponding author

Email address: jpa@ps.au.dk (Jens Peter Andersen)

The concordance between author-based h -indices measured by GS and WoS is largest for Chemistry followed by Cardiology, Political Science, Sociology and Economics.

Keywords: Gender in research, academic careers, Google Scholar, Web of Science, h-index

Highlights

- Examines possible gender differences in the citation coverage of Google Scholar and Web of Science
- Draws on a sample of 1,250 U.S. researchers from five disciplines
- Uses simulations to assess the possible gender consequences of using one database instead of the other
- Finds no notable gender differences in the citation coverage of the two databases
- Simulations indicate marginal effects of database selection on women's success in appointments

1. Introduction

In 2004, weeks after the launch of Scopus, a new, freely accessible citation service, Google Scholar, challenged Web of Science's longstanding monopoly as the sole provider of bibliometric data for citation analysis (Bakkalbasi et al., 2006). Since then, the internal consistency between the databases has attracted much scholarly attention, and for good reasons. Citation metrics have become influential proxies for visibility and scholarly success in academic advancement decisions. Hence, the question of how the choice of citation database influences individual outcomes of bibliometric assessments deserves careful attention. This paper adds to the existing literature by examining whether a shift from Web of Science's "restrictive" citation index to Google Scholar's more "inclusive" citation tracker (Harzing & Mijnhardt, 2015) has any implications for the relative

citation-performance of women and men. A growing literature demonstrates indirect gender effects of how bibliometric indices are used in individual performance assessments (see e.g. Brooks et al., 2014; Nielsen, 2017; Symonds et al., 2006). Yet, we know little about the possible gender bias related to the selection of data sources in such assessments.

Web of Science claims to “*provide access to the most reliable, integrated, multidisciplinary research*”¹. Compared to Google Scholar, the database is characterized by more systematic and transparent criteria for the selection of citation sources and more extensive quality assurance. The database receives its content directly from the journal publishers, and matches extracted citations to specific authors and publications via proprietary algorithms, after standardizing and validating the data (Kulkarni et al., 2009). Google Scholar’s automated citation tracker uses web-crawlers to extract citations from various types of online content². The main criteria for selecting citation sources are that documents should *look scholarly*³ and be publicly available. While WoS describes its proprietary algorithm (DAIS) in some detail⁴, GS’s algorithms remain unspecified. GS is criticized for not specifying its time range and update frequencies, for extracting citations from questionable sources such as power points and funding applications, and for failing to eliminate duplicate sources (Bornmann et al., 2016; Meho & Yang, 2007). In contrast, bibliometricians have raised concerns about WoS’s limited coverage of anthology articles, conference proceedings and monographs, and its bias towards English-language journals (Harzing & van der Wal, 2009).

Studies comparing GS and WoS show large variations in content coverage depending on discipline and time period (Falagas et al., 2007; Neuhaus et al., 2006). Existing comparisons examining the coverage of citing publications, how-

¹<https://apps.webofknowledge.com>

²<https://scholar.google.com/intl/en/scholar/help.html>

³However, Google Scholar also receives some structured content directly from the journal publishers.

⁴<https://clarivate.libguides.com/c.php?g=593069&p=4220414>

ever, typically find that GS captures more unique citations than WoS, especially in the social sciences and humanities (Bosman et al., 2006; Harzing & Alakangas, 2016; Kulkarni et al., 2009; Meho & Yang, 2007; Walters, 2007). This is not surprising given that GS includes many publication types not covered by Web of Science (e.g. doctoral theses, conference proceedings, anthology articles, monographs, regional and online journals, publication outlets in other languages than English, and research reports from policy-oriented think tanks). Some scholars therefore argue that Google Scholar is the most suitable source for measuring the economic and social impacts of scholarly activities (Harzing & van der Wal, 2009).

Research examining how the choice of database influences bibliometric assessments indicates that authors typically garner more citations and higher h-indices in GS than in WoS (Farhadi et al., 2013; Wildgaard, 2015; Minasny et al., 2013; Amara et al., 2013; Mikki, 2010; Franceschet, 2010). However, despite notable differences in coverage, most studies demonstrate good concordance in rankings of individual scholars' citation impact. In a study of 512 authors, Wildgaard (2015) found that h-indices measured by WoS and GS correlated better for researchers in Public Health (Kendall's $\tau = 0.82$), Astronomy (Kendall's $\tau = .79$) and Environmental Science (Kendall's $\tau = 0.79$) than for researchers in Philosophy (Kendall's $\tau = 0.55$). Minasny et al. (2013) computed WoS- and GS-based h-indices for 340 Soil researchers and found very strong correlations (Spearman's $\rho = 0.939$). Amara et al. (2013) demonstrated strong correlations between h-indices measured by WoS and GS for researchers in business and management ($N = 1,286$, Spearman's $\rho = 0.815$) (see also Saad, 2006). De Groote & Raszewski (2012) obtained similar results for nursing researchers ($N = 30$, Pearson's $r = 0.835$). Finally, Franceschet (2010) found a moderate correlation between h-indices measured by GS and WoS for computer scientists (N : unspecified, Kendall's $\tau = 0.52$).

Despite indications of good concordance between author-based citation metrics measured by GS and WoS (with studies of Philosophy and Computer Science as notable exceptions), evidence suggests that choice of database can have cru-

cial implications for the internal ranking of sub-disciplines. Jacobs (2009), for instance, finds that Google Scholar, due to a better coverage of book publications, boosts the citation rates for articles published in the flagship journal for sociology of gender, *Gender & Society*, more than it does for articles in other highly ranked sociology journals. Meho & Yang (2007) compares the scholarly impact of researchers in library and information science (LIS) and find that authors publishing in sub-areas such as communities of practice, computer-mediated communication, data mining, data modeling, discourse analysis and gender and information technology benefit more from having their citation performance measured by Google Scholar than authors publishing in other LIS-related areas.

These findings raise concerns about the possible gender consequences of choice of citation database in individual performance assessments. Existing research demonstrates notable gender differences in primary areas of specialization within disciplines (Dolado et al., 2012; Light, 2013; Elsevier, 2017; Maliniak et al., 2013; West et al., 2013; Andersen et al., 2016). Studies also show that a disproportionate share of women researchers, especially in the social sciences, tend to engage in research topics and methodologies with a lower likelihood of being published in the most prestigious journals (measured by journal impact factors and scholarly rankings) (Dolado et al., 2012; Light, 2013).

Here, we add to this literature by examining to what extent citation data retrieved from WoS and GS reflect the scholarly impact of women and men differently. For this purpose we use a gender-disambiguated sample of 1,250 randomly selected U.S. researchers in Sociology, Political Science, Economics, Cardiology and Chemistry. Our conjecture is that the traditional source, WoS, carries an indirect gender bias in its selection criteria for citation sources that GS avoids due to a more extensive coverage. We also examine the possible gender consequences of using one database instead of the other in simulated hiring scenarios. We calculate database-specific h-indices for all authors in the sample. In repeated simulations, we use these indices to assess whether women’s chances of appointment increase if data are drawn from GS in lieu of WoS.

By examining the possible bias baked into WoS’s “restrictive” citation index,

we contribute important new perspectives to research on gender and bibliometric assessments. This study is also among the first to provide a large-scale, multidisciplinary comparison of the citation coverage of GS and WoS. Most existing research limits its focus to single disciplines and typically relies on fairly small sample sizes (Harzing & van der Wal, 2009), e.g. focusing on a specific academic publication (Bar-Ilan, 2010), a single author (Jacsó, 2008), or a small group of researchers in a selected department or discipline (Meho & Yang, 2007; De Groote & Raszewski, 2012; Franceschet, 2010)⁵. These studies are useful in demonstrating problems with the internal consistency between GS and WoS, but their small sample sizes and monodisciplinary focus make them highly exposed to bias resulting from sample-specific variance. Finally, our study is the first to offer detailed analysis of the coverage of GS and WoS in Sociology, Political Science, Cardiology and Chemistry.

2. Materials & Methods

2.1. Data acquisition

We used an in-house implementation of the WoS citation database, hosted by the Centre for Science and Technology Studies (CWTS), Leiden University. This version allows SQL-based queries and integrates an author-disambiguation algorithm (Caron & van Eck, 2014) particularly well suited for this type of analysis. The algorithm has been found to have very high precision (97%, i.e. few false positives), although missing some publications (recall = 90 – 91%) (Caron & van Eck, 2014). The analysis was based on a random sample of U.S. authors with registered publications in both WoS and GS. We included authors from Sociology, Political Science, Economics, Cardiology and Chemistry. The selected disciplines have different publishing traditions, and we expected them to vary with respect to overlap and internal consistency between GS and WoS records. The following criteria determined the selection of authors:

⁵Notable exceptions in this regard are Wildgaard (2015) and Harzing and Alankangas (2016), but their analyses focus on other disciplinary knowledge domains than ours.

- Eligible authors should have at least ten papers registered in WoS in the period 2008-16
- Eligible authors should be affiliated with research institutions in the United States
- Eligible authors should have at least one paper published in Sociology, Political Science, Economics, Cardiology or Chemistry⁶.

Table 1 shows the total number of eligible authors per discipline.

Table 1: Number of authors identified per included discipline.

Discipline	Number of authors
Political science	4,502
Cardiology	32,522
Economics	11,519
Chemistry	74,150
Sociology	4,389

We developed a simple webform linking eligible authors in WoS to their respective GS profiles. The webform used information on a given author’s first and last name (extracted from WoS) to carry out a semi-automated search query in GS. To ensure field and gender balance in the final sample, author names were represented in a randomized, stochastic order in the webform. Author-specific email addresses and institutional affiliations (extracted from WoS) were used to verify all matches between WoS and GS. In cases of doubt, we looked for resemblances in research interests and publication profiles across WoS and GS. We used this procedure to sample 1,250 eligible authors with active GS profiles - 125 men and 125 women from each discipline.

We used GS-profile IDs to harvest GS publication information (including

⁶We used the following WoS journal categories to identify relevant authors in each discipline: “Sociology”, “Political Science”, “Economics”, “Cardiac & cardiovascular systems” and “Chemistry*”

citation scores and bibliographic metadata) for the 1,250 eligible authors. For this procedure, we used the R-package ‘scholar’ (Keirstead, 2015; R Development Core Team, 2010), resulting in a total of 161,779 records. The metadata extracted from GS is of varying quality and standardization, which complicates the matching to records in WoS. A first iteration of the matching procedure from GS to WoS included obtaining DOI’s through the CrossRef API and reference matching in WoS, however with limited success (Andersen & Nielsen, 2017). We revised this matching procedure, going from WoS to GS instead. We selected only journal articles and reviews ($n = 53,651$) from the included authors, attempting only to match within the individual WoS-GS pairs of author profiles. This limitation allowed us to match for title and journal information, using Levenshtein distance to allow for minor spelling errors (only two edits allowed). The resulting matched set contained 46,082 (85.9%) of the WoS publications. This is a smaller sample than our original setup, however; limiting the set as described above lowers the risk of including GS records falsely assigned to the author in question, and focuses the analysis on the most reliably registered document types. The match corresponds to 28.5% of the GS set, indicating that many of these records are either duplicates, different document types, errors or from non-WoS journals.

The analysis was based on full citation counts as opposed to fixed citation windows. The citation data were originally harvested in May 2017 and updated in June 2018.

2.2. Study setup

To examine gender and disciplinary differences in the citation coverage of GS and WoS, we used standard, descriptive statistics and indicators of coverage, match and ratios of GS citations over WoS citations. Repeated experiments were performed to estimate the possible gender effects of choice of citation database in simulated (indicator-heavy) hiring scenarios (more on this below).

Our key indicator, r , measures the ratio of GS citations over WoS citations. This indicator allows us to estimate how much the citation impact of a paper

is “boosted” if performance is measured by GS in lieu of WoS. An r -value of 2 for instance implies that an individual paper has accrued twice as many citations in GS than in WoS. This indicator is limited to papers with at least one citation in WoS. Means of r , denoted r_{mean} , are calculated as harmonic means. We use r_{mean} to specify the average r value by discipline and gender. An r_{mean} value of 2 for male sociologists implies that men publishing papers in sociology-related journals on average accrue twice as many citations in WoS. We use bootstrapping to estimate the confidence intervals for the harmonic means. Bootstrapping is not ideal for population-like data (Schneider & Van Leeuwen, 2014), but it represents the best possible estimate of uncertainty in r_{mean} scores, especially given the random sample used in this study. To examine the relative citation coverage of GS and WoS we use two indicators: *cover* and *match*. *cover* specifies the mean percentage of references cited by WoS articles to other articles covered by WoS. For instance, a *cover* score of .6 for a male sociologist specifies that 60 percent of the cited references in papers written by male sociologists registered in WoS go to other papers covered by WoS. We expect higher r_{mean} scores in areas with low *cover*, since the potential for GS to cover unique citations is reversely related to the WoS coverage. The *match* indicator specifies the percentage of citations to GS publications that can be matched to WoS publications. A *match* score of .3 for male sociologists would, for instance, imply that on average 30 percent of the papers in the GS profiles of sociologist can be matched to equivalent records in WoS.

To assess the possible influence of co-authorship on the analysis, we examined gender differences in the number of authors per paper by discipline. For this purpose, we recorded per author gender and field the number of co-authors per paper. As displayed in the boxplots in Figure A.4 (Appendix), the women and men included in this sample tend to have largely similar coauthorship frequencies across all five disciplines. Hence, we chose not to adjust for co-authorship frequencies in the analysis.

Finally, we compute database-specific h -indices for all authors in the sample. We denote the h -index h . H -indices were computed in R based on the

matched set of publications only. Irrespective of methodological shortcomings (e.g. Vinkler, 2007; Costas et al., 2010; Van Raan, 2006; Waltman & Van Eck, 2012), the h -index has become an increasingly popular proxy for scholarly performance. This makes it relevant to our analysis. Specifically, we use h -indices to examine whether women’s appointment rates increase if hiring or funding decisions rely on data from GS in lieu of WoS. Existing research documents the use of h -indices in academic hiring and funding decisions in the medical and natural sciences (Barnes, 2014; Marzolla, 2016; Nielsen, 2018; Van den Brink et al., 2013), but we still lack evidence of its general prevalence in such evaluative practices. Further, the use of h -indices tend to be less prevalent in the social sciences than in the natural and health sciences. Hence, our approach should be conceived of as a thought experiment. Our goal is to examine what the possible gender consequences of using GS data instead of WoS data could be, if h -indices were used as the key selection criteria in academic appointment processes.

2.3. Experimental setting

We use experiments to simulate hiring scenarios, where two randomly selected authors (one woman and one man per simulation) are ranked against each other; first, based on WoS h performance, then based on GS h performance. This setup allows us to estimate how the choice of citation database affects women’s chances of obtaining the highest rank in repeated randomized hiring simulations.

We perform repeated simulations for the full sample of 1,250 authors and for each of the five discipline-specific subsamples of 250 authors. To allow for comparisons of scholars at similar career levels, we assigned the authors to age-specific groups and ran full-sample and discipline-specific experiments for each group. The CWTS version of WoS includes information on a given author’s first and last year of publication. In this analysis, we conceptualize an author’s scientific age as the last year of publication minus the first year of publication. We distinguish three age groups in the analysis: junior (scientific age: 0 – 9

years), intermediate (scientific age: 10–19 years), and senior (scientific age: 20+ years). Figure A.3 specifies within-group age variations for women and men in the junior, intermediate and senior group. Naturally, the largest age variations occur in the senior group, which does not have a pre-defined upper bound. There are also notable gender and disciplinary differences for the junior group, with somewhat lower median ages in Chemistry and Cardiology, especially for women. Some disciplines in the sample include no authors under the scientific age of six. Note here that authors must have published at least ten papers to be included in the analysis. In other words, the age distributions for the junior group indicate that it takes longer to author ten papers in economics than in chemistry.

All experiments were performed with $n_{male} + n_{female}$ random samples. For some of the models, this is a quite small sample, and we should expect this to be reflected in the results.

3. Results

3.1. Descriptive data

Figure 1 displays gender and disciplinary differences in mean-citation ratios (i.e. r_{mean}) per paper as a function of the percentage of papers covered by WoS (i.e. $cover$). Differences in r_{mean} are given on the y -axis and differences in $cover$ on the x -axis (see Table A1 for numerical specifications). The r_{mean} ratios and confidence intervals indicate inconsistent gender variations across the five disciplines. Publications by women have notably higher r_{mean} ratios in Sociology ($f = 2.08$ vs. $m = 1.73$), Cardiology ($f = 1.49$ vs. $m = 1.19$) and Political Science ($f = 2.54$ vs. $m = 1.80$), while publications by men score marginally higher in Chemistry ($f = 1.19$ vs. $m = 1.32$), and women marginally higher in Economics ($f = 1.99$ vs. $m = 1.87$). However, except for Political Science, the confidence intervals are too broad to consider any of these differences consequential.

Figure 1 also demonstrates a tendency towards smaller citation ratios (and narrower confidence intervals) as the *cover* indicator increases along the *x*-axis. In Political Science, Sociology and Economics, where the WoS coverage is low (45% – 63%), the r_{mean} ratios are relatively high, whereas the opposite is true for Cardiology and Chemistry, where WoS coverage is high (77% – 84%). The match between publications in GS and WoS is relative low in all disciplines (18% – 36%) (See Table 4 and Figure A.1 for numerical specifications), which demonstrates the importance of integrating citations from both databases in bibliometric assessments of individual researchers’ performance.

Figure 1: r_{mean} and r_{ci} per paper as function of coverage per discipline and gender.

Since we focus on the consequences for authors, we also report the ratios of mean citations per author in Figure 2. These are based on almost the same data as the per-paper ratios. However, 34 observations were removed due to an extraordinarily low matching rate between WoS and GS. It is most likely that the persons behind the matched profiles were in fact not identical. As this is of little consequence above (the included papers are still matched, the few included papers are found due to co-authorships), it matters for the per-author analysis. In all the following per-author analyses, these observations are therefore excluded. The main difference in the results shown in Figures 1 and 2 is due

to the author means dampening the extreme r -values of individual papers. The discipline-pattern is the same, but the gender variation changes for the r_{mean} -scores. Women now have lower scores than men in Sociology, Political Science and Economic (marginally), while slightly higher in Cardiology and Chemistry. All confidence intervals for gender-pairs now overlap. Interestingly, the ratios also increase for the means, which shows how much low-cited publications influence the bootstrap confidence intervals in the per-paper ratios.

Figure 2: r_{mean} and r_{ci} per author, grouped by discipline and gender.

3.2. Academic appointment simulation

Before we present the results of the simulated hiring scenarios, we briefly examine the concordance between per-author h -indices measured by GS and WoS for each discipline. For this purpose, we use Spearman's ρ and Kendall's τ rank-correlation coefficients. As displayed in Table 2, Spearman's ρ is very strong across all disciplines (.92 – .99), while Kendall's τ is higher in Chemistry and Cardiology (.93 – .94) than the other fields (.80 – .87). The ρ values are higher than expected in Economics, Political Science and Sociology, as previous studies indicate higher levels of concordance between the citation databases in the natural and health sciences and in disciplines with good WoS coverage

(Bosman et al., 2006; Harzing & Alakangas, 2016; Kulkarni et al., 2009; Meho & Yang, 2007; Walters, 2007). The τ values capture this difference more clearly.

Table 2: Rank correlation coefficients of WoS and GS h -indices per discipline.

Discipline	Spearman’s ρ	Kendall’s τ
Chemistry	0.99	0.94
Cardiology	0.99	0.93
Economics	0.92	0.80
Political Science	0.96	0.87
Sociology	0.96	0.87

In the following we report the results of the appointment simulation, for the full set of authors in each discipline (general model), the three academic age groups specified above, as well as a sample using the authors in the WoS h -interquartile range (distribution model). It should be noted that the h -indices used in this simulation are not those given by full WoS and GS profiles, but recalculated based on the matched set of publications found in this study. In a real setting, an applicant would use the full publication set, however, as we test the citation difference and not the publication difference, this is not relevant here.

Figure 3 reports the outcomes of the simulated hiring scenarios. It specifies female authors’ odds ratios of outperforming their male competitors in pairwise simulations, when h -indices are measured by GS in lieu of WoS. The results presented in Figure 3 are calculated based on $n_{male} + n_{female}$ random samples per discipline and age group. Here, we specify the underlying calculation of this procedure. We use the outcomes of the senior model for Cardiology as example (Table 3). The upper left cell (a) specifies the number female hires when citation data are drawn from GS. The upper right cell (b) reports the number of successful females with data drawn from WoS. The lower left cell (c) reports the number of successful males using GS data; and the lower right cell (b) specifies that number of successful male candidates based on WoS data. The

standard odds-ratio (OR) formula and the calculation of the odds-ratio value, standard error (SE) and 95% confidence interval (CI) are shown below. The interpretation of the odds-ratio is the increased/decreased likelihood of selecting a woman, when shifting from the WoS h -index to the GS h -index in simulated hiring decisions.

Table 3: Absolute values of the contingency table corresponding to the “Senior Cardiology” experiment.

	GS	WoS
Female	18(a)	16(b)
Male	23(c)	25(d)

$$\begin{aligned}
OR &= \frac{a/b}{c/d} = \frac{18/16}{23/25} = 1.22 \\
SE &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{18} + \frac{1}{16} + \frac{1}{23} + \frac{1}{25}} = .45 \\
CI &= e^{\ln(OR) - 1.96SE}, e^{\ln(OR) + 1.96SE} \\
&= e^{\ln(1.22) - 1.96*.45}, e^{\ln(1.22) + 1.96*.45} = .51; 2.95
\end{aligned}$$

The confidence intervals in Figure 3 are quite broad, regardless of the model. The odds-ratios range from no effect to large effects. However, all model combinations have confidence intervals spanning the line of no difference (i.e. Odds ratio = 1), meaning we can not reliably conclude on any differences. The numerical values for the odds-ratios and 95% confidence intervals can be found in table 4. Within each model we only find small odds-ratios, but when combining these in the general model, the tendencies for Political Science and Sociology become stronger, but still with the lower bounds of the confidence intervals at .68 and .62 respectively. In summary, our sample reveals no difference in the selected gender, regardless of which database the h -indices are calculated from.

Figure 3: Odds-ratio and corresponding confidence interval for selecting a female candidate using GS h instead of WoS h . Each result shows the number of male (m) and female (f) candidates in the sampling pool. Sampling is repeated at $(n_m + n_f)$ iterations.

Table 4: For each discipline and model, odds-ratios and 95% confidence intervals.

	Junior	Intermediate	Senior	Distribution	General
Cardiology	.9 (.36-2.25)	1 (.42-2.39)	1.22 (.51-2.95)	1 (.42-2.38)	1.22 (.51-2.89)
Chemistry	1.22 (.51-2.95)	1 (.42-2.38)	1 (.42-2.39)	1 (.29-3.4)	1 (.42-2.41)
Economics	1.22 (.51-2.92)	1.22 (.51-2.95)	.82 (.35-1.96)	.9 (.37-2.18)	.9 (.37-2.18)
Political Science	1.13 (.43-2.96)	1.23 (.5-2.99)	1.36 (.56-3.29)	1 (.42-2.39)	1.64 (.68-3.94)
Sociology	.91 (.38-2.17)	1.24 (.5-3.1)	.91 (.38-2.17)	1.22 (.51-2.95)	1.51 (.62-3.69)

4. Concluding discussion

This study set out to examine possible gender differences in the citation coverage of GS and WoS in five disciplines. We also investigated the potential gender consequences of using one database instead of the other in simulated hiring scenarios. In general, the analyses indicate mostly marginal and inconsistent gender effects. The descriptive analysis of author-based r_{mean} scores (i.e. the average “citation boost” of measuring impact by GS as opposed to WoS) (Figure 2) reveals close to identical results for women and men across disciplines. In the simulated-hiring scenarios, we observe no gender effects, regardless of discipline and age group. While there is a potential difference in favor of women using GS in the general model for Political Science and Sociology, the confidence intervals are too wide to consider these gender variations consequential (Figure 3). A larger sample size may show a clearer trend. Yet, based on the results of this study, the choice of database appears to be of little influence on women’s and men’s relative performance and appointment success. A key limitation of the study lies in its focus on researchers from the US. Thus, it remains to be explored whether these results hold for researchers from other countries.

Another limitation concerns the basic presumption that guide our simulated hiring experiment: that decision-makers use h -indices to select between candidates in individual performance assessments. While a number of studies document the use of h -indices in academic hiring and funding decisions (Barnes, 2014; Marzolla, 2016; Nielsen, 2018; Van den Brink et al., 2013), we still lack evidence on how widespread the evaluative use of h -indices may be in funding and hiring decisions.

In line with the existing literature demonstrating relatively low coverage of the social science literature in WoS, we find that the mean percentage of citations in WoS papers that refer to other articles covered by WoS is smaller in the social sciences than in the health and natural sciences. The concordance between rankings of h -indices measured by GS and WoS is also higher in Chemistry and Cardiology than in Economics, Political Science and Sociology. Consequentially,

the possible assessment bias related to choice of database is largest in the social sciences.

Evaluators should be attentive to such inconsistencies and carefully consider how their choice of data source influences assessments at the individual level. This is particularly pertinent in the social sciences, where a combination of citation databases may be preferable to a single source. In general, citation indices should be used with extreme caution and only as supporting information for other assessment types (Vinkler, 2007; Costas et al., 2010; Aksnes & Taxt, 2004).

Acknowledgements

We wish to extend our gratitude for fruitful feedback and questions from the audience at the ISSI conference 2017, where an early version of this work was presented. We would also like to thank two anonymous reviewers for truly constructive feedback.

References

References

- Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation*, 13, 33–41. doi:10.3152/147154404781776563.
- Amara, N., Landry, R., & Halilem, N. (2013). Faculty consulting in natural sciences and engineering: Between formal and informal knowledge transfer. *Higher Education*, 65, 359–384. doi:10.1007/s10734-012-9549-9.
- Andersen, J. P., & Nielsen, M. W. (2017). Do Google Scholar and Web of Science reflect women’s and men’s scholarly impact differently? A comparison of U.S. researchers in sociology and economics. In *Proceedings of the 16th International Conference on Scientometrics and Informetrics*. Wuhan, China.

- Andersen, J. P., Schneider, J. W., & Nielsen, M. W. (2016). Mapping the author gender-distribution of disease-specific medical research. In *Proceedings of the 21st International Conference on Science and Technology Indicators* (pp. 852–862). Valencia, Spain.
- Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3, 1–8. doi:10.1186/1742-5581-3-7.
- Bar-Ilan, J. (2010). Google Scholar Citations to Information and Library Science Journals. In *Eleventh International Conference on Science and Technology Indicators Book of Abstracts* (pp. 28–30). Leiden, Netherlands.
- Barnes, C. (2014). The emperor’s new clothes: the h-index as a guide to resource allocation in higher education. *Journal of Higher Education Policy and Management*, 36, 456–470. doi:10.1080/1360080X.2014.936087.
- Bornmann, L., Thor, A., Marx, W., & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology*, 67, 2778–2789. doi:10.1002/asi.23627.
- Bosman, J., van Mourik, I., Rasch, M., Sieverts, E., & Verhoeff, H. (2006). *Scopus reviewed and compared*. Technical Report Utrecht University Library Utrecht, Netherlands.
- Brooks, C., Fenton, E. M., & Walker, J. T. (2014). Gender and the evaluation of research. *Research Policy*, 43, 990–1001. doi:10.1016/j.respol.2013.12.005.
- Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the Science and Technology Indicators conference 2014* (pp. 79–86). Leiden, Netherlands: CWTS, Leiden University.

- Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61, 1564–1581. doi:10.1002/asi.21348.
- De Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing. *Nursing Outlook*, 60, 391–400. doi:10.1016/j.outlook.2012.04.007.
- Dolado, J. J., Felgueroso, F., & Almunia, M. (2012). Are men and women-economists evenly distributed across research fields? Some new empirical evidence. *SERIEs*, 3, 367–393. doi:10.1007/s13209-011-0065-4.
- Elsevier (2017). *Gender in the Global Research Landscape*. Technical Report Elsevier. doi:10.17632/bb3cjfgm2w.2.
- Falagas, M. E., Pitsouni, E. I., Malietzis, G. a., & Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22, 338–342. doi:10.1096/fj.07-9492LSF.
- Farhadi, H., Salehi, H., Yunus, M. M., Chadegani, A. A., Farhadi, M., Fooladi, M., & Ebrahim, N. A. (2013). Does it Matter Which Citation Tool is Used to Compare the h-index of a Group of Highly Cited Researchers? *Australian Journal of Basic and Applied Sciences*, 7, 198–202.
- Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83, 243–258. doi:10.1007/s11192-009-0021-2.
- Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106, 787–804. doi:10.1007/s11192-015-1798-9.

- Harzing, A.-W., & Mijnhardt, W. (2015). Proof over promise: towards a more inclusive ranking of Dutch academics in Economics & Business. *Scientometrics*, 102, 727–749. doi:10.1007/s11192-014-1370-z.
- Harzing, A.-W., & van der Wal, R. (2009). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science and Technology*, 60, 41–46. doi:10.1002/asi.20953.
- Jacobs, J. A. (2009). Where credit is due: Assessing the visibility of articles published in Gender & Society with Google Scholar. *Gender and Society*, 23, 817–832. doi:10.1177/0891243209351029.
- Jacsó, P. (2008). Google Scholar revisited. *Online Information Review*, 32, 102–114. doi:10.1108/14684520810866010.
- Keirstead, J. (2015). *scholar: analyse citation data from Google Scholar*. R package version 0.1.4.
- Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of Citations in Web of Science ,. *Jama*, 302, 1092–1096.
- Light, R. (2013). Gender Inequality and the Structure of Occupational Identity: The Case of Elite Sociological Publication. In S. McDonald (Ed.), *Networks, Work and Inequality (Research in the Sociology of Work, Volume 24)* (pp. 239–268). Emerald Group Publishing Ltd.
- Maliniak, D., Powers, R., & Walter, B. F. (2013). The Gender Citation Gap in International Relations. *International Organization*, 67, 889–922. doi:10.1017/S0020818313000209.
- Marzolla, M. (2016). Assessing evaluation procedures for individual researchers: The case of the Italian National Scientific Qualification. *Journal of Informetrics*, 10, 408–438. doi:10.1016/j.joi.2016.01.009.

- Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, . doi:10.1002/asi.20677.
- Mikki, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, *82*, 321–331. doi:10.1007/s11192-009-0038-6.
- Minasny, B., Hartemink, A. E., McBratney, A., & Jang, H.-J. (2013). Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ*, *1*, e183. doi:10.7717/peerj.183.
- Neuhaus, C., Neuhaus, E., Asher, A., & Wrede, C. (2006). The Depth and Breadth of Google Scholar: An Empirical Study. *portal: Libraries and the Academy*, *6*, 127–141. doi:10.1353/pla.2006.0026.
- Nielsen, M. W. (2017). Gender consequences of a national performance-based funding model: new pieces in an old puzzle. *Studies in Higher Education*, *42*, 1033–1055. doi:10.1080/03075079.2015.1075197.
- Nielsen, M. W. (2018). Scientific Performance Assessments Through a Gender Lens. *Science & Technology Studies*, *31*, 2–30. doi:10.23987/sts.60610.
- R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Saad, G. (2006). Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. *Scientometrics*, *69*, 117–120. doi:10.1007/s11192-006-0142-9.
- Schneider, J. W., & Van Leeuwen, T. N. (2014). Analysing robustness and uncertainty levels of bibliometric performance statistics supporting science policy. A case study evaluating Danish postdoctoral funding. *Research Evaluation*, *23*, 285–297. doi:10.1093/reseval/rvu016.

- Symonds, M. R., Gemmell, N. J., Braisher, T. L., Gorringer, K. L., & Elgar, M. A. (2006). Gender differences in publication output: Towards an unbiased metric of research performance. *PLoS ONE*, *1*, 1–5. doi:10.1371/journal.pone.0000127.
- Van den Brink, M., Fruytier, B., & Thunnissen, M. (2013). Talent management in academia: Performance systems and HRM policies. *Human Resource Management Journal*, . doi:10.1111/j.1748-8583.2012.00196.x.
- Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*, 491–502. doi:10.1556/Scient.67.2006.3.10.
- Vinkler, P. (2007). Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of Information Science*, *33*, 481–491. doi:10.1177/0165551506072165.
- Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing and Management*, *43*, 1121–1132. doi:10.1016/j.ipm.2006.08.006.
- Waltman, L., & Van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, *63*, 406–415. doi:10.1002/asi.21678.
- West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The Role of Gender in Scholarly Authorship. *PLoS ONE*, *8*. doi:10.1371/journal.pone.0066212.
- Wildgaard, L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar. *Scientometrics*, *104*, 873–906. doi:10.1007/s11192-015-1608-4.

A. Appendix

Figure A.1: Percentage of matched records as function of coverage for discipline and gender.

Figure A.2: r_{mean} and r_{ci} per author, grouped by discipline and gender. This figure includes the single outlying observation excluded in the main analysis.

Table A.1: r_{mean} , r_{ci} , $cover$ and $match$ values per discipline and gender.

Discipline	Gender	r_{mean} (r_{ci})	$cover$	$match$
Sociology	Men	1.73 (1.21-2.4)	.521	.228
Sociology	Women	2.08 (1.83-2.34)	.601	.286
Political science	Men	1.80 (1.49-2.19)	.448	.183
Political science	Women	2.54 (2.28-2.78)	.498	.205
Economics	Men	1.87 (1.56-2.2)	.575	.210
Economics	Women	1.99 (1.55-2.41)	.628	.229
Cardiology	Men	1.19 (.93-1.46)	.835	.313
Cardiology	Women	1.49 (1.37-1.6)	.825	.331
Chemistry	Men	1.32 (1.2-1.42)	.773	.356
Chemistry	Women	1.19 (.94-1.42)	.825	.362

Figure A.3: Distribution of authors per scientific age category, gender and discipline.

Figure A.4: Distribution of “number of authors per paper”, per gender and discipline.