



# Quantitative analysis of automatic performance evaluation systems based on the *h*-index

Marc P. Hauer<sup>1</sup> · Xavier C. R. Hofmann<sup>1</sup> · Tobias D. Krafft<sup>1</sup> · Katharina A. Zweig<sup>1</sup>

Received: 18 April 2019 / Published online: 14 March 2020  
© The Author(s) 2020

## Abstract

Since the *h*-index has been invented, it is the most frequently discussed bibliometric value and one of the most commonly used metrics to quantify a researcher's scientific output. The more it is increasingly gaining popularity to use the metric as an indication of the quality of a job applicant or an employee the more important it is to assure its correctness. Many platforms offer the *h*-index of a scientist as a service, sometimes without the explicit knowledge of the respective person. In this article we show that looking up the *h*-index for a researcher on the five most commonly used platforms, namely AMiner, Google Scholar, ResearchGate, Scopus and Web of Science, results in a variance that is in many cases as large as the average value. This is due to the varying definitions of what a scientific article is, the underlying data basis, and different qualities of the entity recognition problem. To perform our study, we crawled the *h*-index of the worlds top researchers according to two different rankings, all the Nobel Prize laureates except Literature and Peace, and the teaching staff of the computer science department of the TU Kaiserslautern Germany with whom we additionally computed their *h*-index manually. Thus we showed that the individual *h*-indices differ to an alarming extent between the platforms. We observed that researchers with an extraordinary high *h*-index and researchers with an index appropriate to the scientific career path and the respective scientific field are affected alike by these problems.

**Keywords** Bibliometrics · Big data · *h*-index

## Introduction

Using online services to get more information about the quality of a job applicant or an employee is increasingly gaining popularity, as information gathered through platforms like, for example, Facebook or LinkedIn can be automatically processed without the need of investing personnel expenditure (Sattelberger 2015). Furthermore, the process

---

✉ Marc P. Hauer  
hauer@cs.uni-kl.de

<sup>1</sup> Algorithm Accountability Lab, Department of Computer Science, Technische Universität Kaiserslautern, Gottlieb-Daimler-Straße 48, 67663 Kaiserslautern, Germany

suggests neutrality and fairness (Gapski 2015), implying an objective treatment of the individual, since every person is evaluated by the same success measures. Using a bibliometric measure to evaluate the performance of scientists, for example, is common practice (Nature 2017). One of the most frequently used of such metrics is the  $h$ -index (Ball 2007; Saleem 2011). It computes a value depending on the number of papers published by a scientist and their respective impact on other researches such that:

A scientist has index  $h$ , if  $h$  of his or her  $N_p$  papers have at least  $h$  citations each and the other  $(N_p - h)$  papers have  $\leq h$  citations each (Hirsch 2005).

The validity of such a one-dimensional evaluation of scientists relies heavily on a complete list of all publications the scientist authored, as well as all publications which cite any of those. Since manually gathering such a list is impossible, the utilization of citation databases is inevitable. The most popular platforms offering the  $h$ -index are: AMiner (2017a), Google Scholar (2017), ResearchGate (2017), Scopus (2017) and Web of Science (2017). Searching for the  $h$ -index of multiple names on all of those platforms it can be seen that a number of inconsistencies between the different platforms like wrong assignments of works to an author (entity resolution errors), missing, doubled or false database entries and many more sources of error (MacRoberts and MacRoberts 1989) can lead to huge discrepancies in the resulting  $h$ -indices (and make it almost impossible to correctly estimate the exact value). Additionally, it is generally known that citation behaviour often varies greatly between different scientific disciplines. Various studies have already proven this, but the investigations are either directed at a single platform where the names and associated scientific disciplines can be extracted from the citation database itself (Batista et al. 2006), which limits the validity of the results due to the error-proneness of such an approach, or to multiple platforms with a very small set of names (Bar-Ilan 2008).

For promotion committees at a university looking up the  $h$ -index value of a potential employee on at least one of the platforms, it is important to know to what extent those errors may influence the results. Therefore, we compare the aforementioned five platforms offering the computation of the  $h$ -index, following four big research questions:

RQ1 (Differences between platforms): To what extent are there differences of  $h$ -index value distributions for a given sample between the platforms?

RQ2 (Difference between values for the individuals): For a given set of persons, how big are the individual discrepancies of  $h$ -index values?

RQ3 (Differences between scientific disciplines): How much do the  $h$ -indices for Nobel Prize winners differ depending on their scientific discipline on the various platforms?

RQ4 (Comparison to ground truth): Of what magnitude are the differences compared to the persons manually assessed  $h$ -indices?

To answer these questions, we first introduce the platforms in “[The platforms](#)” section. We then present the results of our study for which we gathered the  $h$ -indices on said platforms for different sets of names and determine the aggregated gaps in “[Study I](#)” section. Afterwards we evaluate how these differences are reflected in the  $h$ -indices of the nobel prize winners when we differentiate by scientific discipline (“[Study II](#)” section). Finally, we compute the reasonably accurate  $h$ -indices for a test group of 25 names by hand in order to compare them to the results delivered by the platforms and to inspect the respective deviations, which is presented in “[Study III](#)” section. In “[Threats to validity](#)” section, we will analyze potential threats to validity and in “[General discussion](#)” section, we will conclude the results in a general discussion.

## The platforms

In this section, we discuss the platforms used in our studies and their individual properties. In Table 1 those properties are summarized and extended by additional common aspects. We are especially interested in the question of whether scientists create their own accounts or whether an account can also be automatically created by the system or by someone else.

*AMiner (AM)* started as a research project led by Dr. Jie Tang at Tsinghua University, China. Based on the comparably lower number of unique surnames in China than in other countries, one of its primal goals is to differentiate between multiple people with the same name, but at the same time it ignores the fact that a person may publish under several name variations (e.g. different name abbreviations) (Tang et al. 2008). The database is constructed by crawling a variety of different web-sources, which leads to automatically constructed profiles without a permission or notice. This procedure and additional manually constructed profiles may lead to duplicated and non-scientific entries. The platform also allows the user to manually correct mistakes and complete profiles. Manual modification does not require any form of authentication or validation and therefore allows easy manipulation of profiles.

*Google Scholar (GS)* is a free to use platform for scientific investigation provided by Alphabet. It screens websites for a certain kind of formatting and checks the indexing of online documents to decide whether they are scientific publications. Due to the susceptibility to errors of this way of data extraction, many unscientific contributions are listed on GS (Petersen et al. 2014). A profile has to be constructed manually, automatically constructed profiles are created only for deceased people such as Sigmund Freud or Albert Einstein.

*ResearchGate (RG)* is a social network for researchers and scientists, focusing on the person as central entity instead of their work. In general, accounts can be manually created by a scientist or automatically constructed. Only on manually created accounts, the *h*-index is shown. If such an account is abandoned later, the indexing algorithm automatically adds further papers and increases the *h*-index accordingly, though the respective person can configure a mandatory manual validation.

*Scopus (SP)* is a platform that offers a number of services which are strongly limited depending on whether the user has a purchased or free access. Since the free version provides all functionality important for this study, the following description focuses on the latter. The database is constructed by extraction of bibliometric information from a specific set of journals (on the platform this set is called content coverage) this set is publicly visible. In case of missing publications or mistakes, the automatically constructed profiles cannot be edited by the user, a time-consuming support system has to be used instead. Whether and to what extent user feedback will actually influence the database is unclear.

*Web of Science (WoS)* is a database set up in a similar way as SP, by screening publications from a limited set of journals. It has to be noted that the platform does not support profiles of any kind to provide bibliometric information of authors. Instead, it allows the dynamic construction of a so-called citation report, which contains any publication released by a person with the name searched for. From this report, falsely assigned publications can be excluded manually to correct the displayed *h*-index. This process leads to a comparatively low error rate. However, the strongly limited selection of journals considered results in rather low values (Piwowar 2013; Nature 1965).

All of the presented platforms lack transparency regarding the limited validity of the *h*-indices they provide. Even well-known basic aspects that have to be considered when it comes to the *h*-index, for example the incomparability of values between different scientific

**Table 1** Important aspects of the platforms

	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Charged	No	No	No	Yes	Yes
Primary entity	Scientists	Publications	Scientists and publications	Publications	Publications
Integration of OreID	No	No	No	Yes	Yes
Primary datasource	Manual uploads, webcrawling	Webcrawling	Webcrawling	Manual journal selection by the maintainer	Manual journal selection by the maintainer
Advice about how to use the <i>h</i> -index	No	No	No	No	No
Account creation	Manually, automatically	Manually, (rarely) automatically	(Primary) automatically, manually	Automatically	No accounts, dynamic profiles through article processing (article groups)
Account administration	Manually	Manually	Manually (universal access)	Upon request	Upon request
Multiple accounts per person	Can be found	Not found	Can be found	Can be found	Can be found in article groups
Falsely assigned authorships	Can be found	Can be found	Can be found	Can be found	Can be found in article groups
Multiple entities of the same publication	Can be found according to the manual	Can be found	Can be found	Not found	Not found
Unscientific entities	Can be found	Can be found	Can be found	Can be found	Not found

fields, are nowhere to be found, let alone explanations about less-known potential sources of errors. Though all of them list an explanation about how they gathered their data, they do not note that they cannot guarantee for complete publication coverage. Additionally, none of the platforms seems to consider a differentiation between scholarly articles and others beyond the method they use for crawling the data, at least there is no information available indicating such a differentiation. Last but not least the crucial information to what extent authors actually administrate their profile is not visible. Thus, we were interested in how much the  $h$ -indices would differ for a wide range of scientists. In the following, we will describe three different studies consisting of various lists of scientists for which we gathered their  $h$ -index, and the resulting variance of a person's  $h$ -index over the different platforms.

## Study I

### Description

In order to find out the differences between the platforms as posed in RQ1 we collected several sets of names and developed a scraper visiting each platform once per name, extracting the respective  $h$ -index and saving it to a database for elaboration. Subsequently, the maximum deviation of the average  $h$ -index of individuals compared to the  $h$ -index found on the platforms was analyzed in order to address RQ2.

### Method

*Step 1* Setup sets of names: To encounter multiple possible threats to validity we chose four sets of names with different criteria to analyze as described in the following.

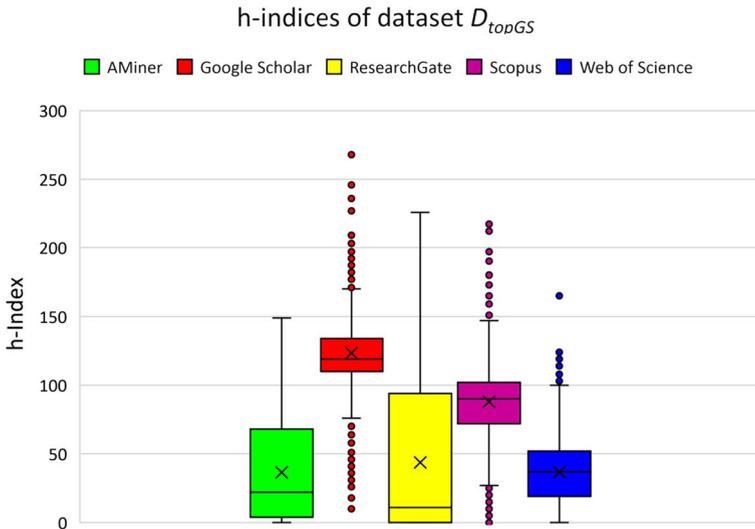
- $D_{\text{topGS}}$  contains 1360 names of the researchers with an  $h$ -index of at least 100 according to the Webometrics Ranking of World Universities initiative of the Cybermetrics Lab research group in Spain (Webometrics 2017), who get their data from GS. Since some of the names are only available in an abbreviated form, we removed them to increase the chance of correct evaluation. After doing so, 1295 names were left.
- $D_{\text{topAM}}$  contains 139 names of the researchers with an  $h$ -index of at least 100 according to AM (2017b).
- $D_{\text{nobel}}$  contains the 632 names of the researchers who won a Nobel Prize in chemistry, physics, medicine or economics (Nobel 2017).
- $D_{\text{TUKL}}$  contains the 56 names of the current docents of a technical department at a German university.

*Step 2* Collect  $h$ -indices: Most of the platforms do not offer an API to allow easy automatic access. As a consequence, the scraper visits the respective websites via browser by using the Selenium testing environment (Razak and Fahrurazi 2011) to search each name from a sequentially loaded subset of names and extracts the returning  $h$ -indices for the first result found respectively, to simulate an employers behavior. The  $h$ -indices are stored in an Elasticsearch database, which allows easy access, modification and visualization via Kibana.

*Step 3* Refine the results: Since the crawling procedure is prone to failure due to occasional loading errors, delays and further problems beyond our control, some subsets

**Table 2** The number of names in each dataset, for which an  $h$ -index on all platforms could be found

$D_{\text{topGS}}$	$D_{\text{topAM}}$	$D_{\text{topNobel}}$	$D_{\text{TUCL}}$
869 (of 1360)	62 (of 139)	89 (of 632)	32 (of 56)

**Fig. 1** Box plots of the  $h$ -indices for the names in  $D_{\text{topGS}}$ 

delivered very little results. To avoid a lack of applicable data those subsets were reprocessed. Names with accented or special characters that might not be processed correctly have been excluded from the study as well.<sup>1</sup>

*Step 4 Evaluation:* Sometimes, a name cannot be found on all platforms discussed. Therefore we restrict the database for our analysis to those names which were found on all platforms (see Table 2). The found  $h$ -indices are split by platform and visualized as box plots for each dataset, respectively.

## Results

The box plots in Fig. 1 unveil notable differences between the platforms for each set of names. The key values for  $D_{\text{topGS}}$  as presented in Table 3 show that the average  $h$ -index differs up to a factor of 3 between AM and GS, the median between RG and GS even by a factor of 12. The Inter Quartile Ranges (IQR) for the values from AM and RG are at least twice as high of those for GS, SP and WOS (see Fig. 1 and Table 3).

Except for AM, the plots for  $D_{\text{topAM}}$  look very similar (see Fig. 2 and Table 4).

<sup>1</sup> This measure affects only the following 10 names: Jean Pierre Després, John O’Keefe, José Ordovás, József Pálkás, Mark D’Esposito, Paul Erdős, Robert Bárány, Santiago Ramón y Cajal, Vincent Lamaître, Zoltán Trócsányi.

**Table 3** Key values describing the results of  $D_{topGS}$

$D_{topGS}$	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Average	36.543	123.354	43.779	87.918	36.763
Median	22	119	11	90	37
Upper quartile	68	134	94	102	52
Lower quartile	4	110	0	72	19
IQR	64	24	94	30	33
Upper antenna	149	170	226	147	100
Lower antenna	0	76	0	27	0

**Table 4** Key values describing the results of  $D_{topAM}$

$D_{topAM}$	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Average	114.71	146.623	50.952	110.5	48.323
Median	111	141.5	9.5	107.5	48.5
Upper quartile	121	168.75	101.75	138.5	67.5
Lower quartile	105.75	123	0	78.75	28.75
IQR	15.25	45.75	101.75	59.75	38.75
Upper antenna	143	236	212	211	97
Lower antenna	100	88	0	49	6

**Table 5** Key values describing the results of  $D_{nobel}$

$D_{nobel}$	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Average	11.696	57.293	13.663	45.772	15.391
Median	5	45.5	3	39.5	12
Upper quartile	13	89	15.75	63	24
Lower quartile	2	21.25	0	15.75	2.25
IQR	11	57.75	15.75	47.25	21.75
Upper antenna	29	188	38	133	53
Lower antenna	0	0	0	0	0

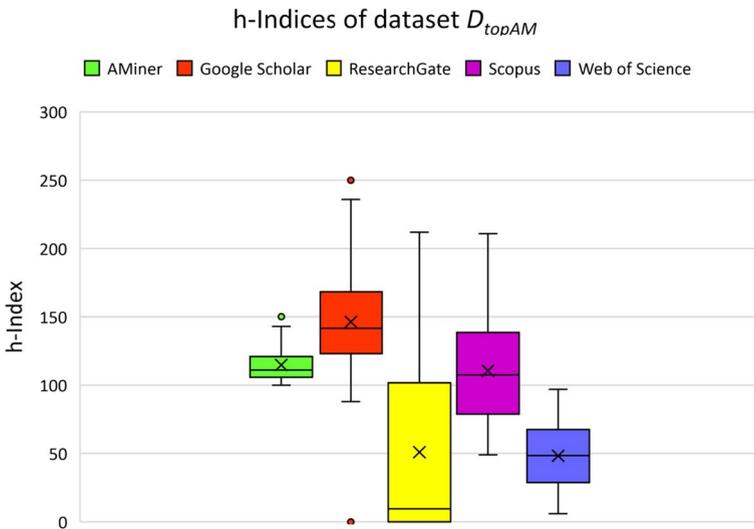
For  $D_{nobel}$ , the distribution between the plots look much like those for  $D_{topGS}$  except that they are lower (see Fig. 3 and Table 5). This could be explained by the different criteria for the selection of scientists (citation based on  $D_{topGS}$  vs. Nobel Prize based on  $D_{nobel}$ ).

$D_{TUKL}$  focuses on scientists from only one department of a university. Accordingly, the resulting values are lower and lead to different distributions. While  $h$ -indices from GS are still greater than the ones from other platforms, AM, RG and SP yield similar values compared to each other (see Fig. 4 and Table 6).

To inspect the worst-case impact of the findings for individuals, we examined the maximum and minimum  $h$ -indices from any of the platforms for each of the 1052 names in all

**Table 6** Key values describing the results of  $D_{TUKL}$ 

$D_{TUKL}$	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Average	10.625	18.375	11.594	12.469	6.938
Median	9.5	15.5	11	9.5	5.5
Upper quartile	13	27	16.75	15.75	9.75
Lower quartile	3.25	7.5	5	5	3.25
IQR	9.75	19.5	11.75	10.75	6.5
Upper antenna	25	46	27	29	17
Lower antenna	1	3	0	3	1

**Fig. 2** Box plots of the  $h$ -indices for the names in  $D_{topAM}$ 

datasets sorted by their average  $h$ -index on all five platforms (see Fig. 5). It turns out that the discrepancies between maximum and minimum are considerably high.

## Discussion

Scientific evidence that different platforms yield different results when it comes to the  $h$ -index values and therefore address RQ1 has already been provided by others (Falagas et al. 2008). Consequently, we did not expect to find very similar  $h$ -index values, however, the extent of differences was considerably higher than expected. Especially the  $h$ -index values from GS generally seem to be higher than on the other platforms. The values found for  $D_{topAM}$  are approximately as high for AM as for RG. This is due to the fact that the dataset focuses on names with a high  $h$ -index on AM itself. WOS is considered a well-known and well-used tool for scientific literature research, but due to its relatively small coverage of journals (Reuters 2008), the gathered  $h$ -indices are comparably low.

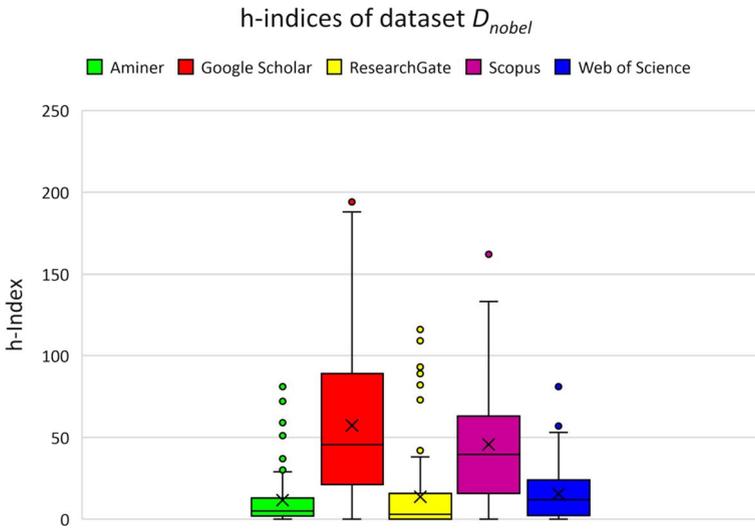


Fig. 3 Box plots of the  $h$ -indices for the names in  $D_{nobel}$

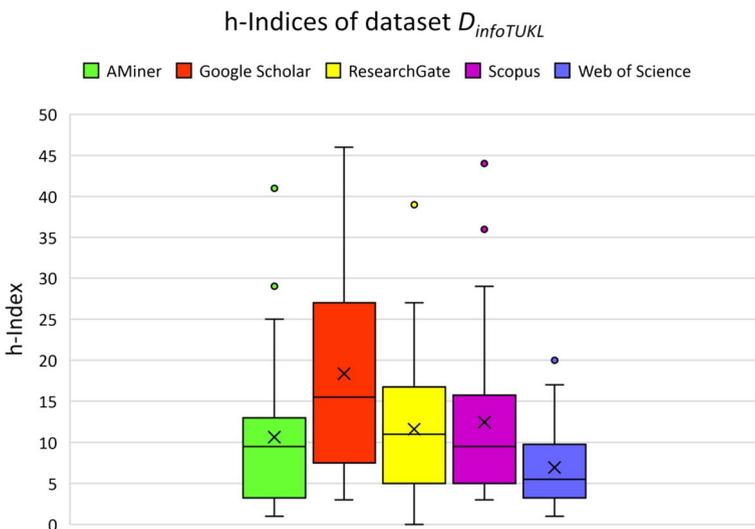
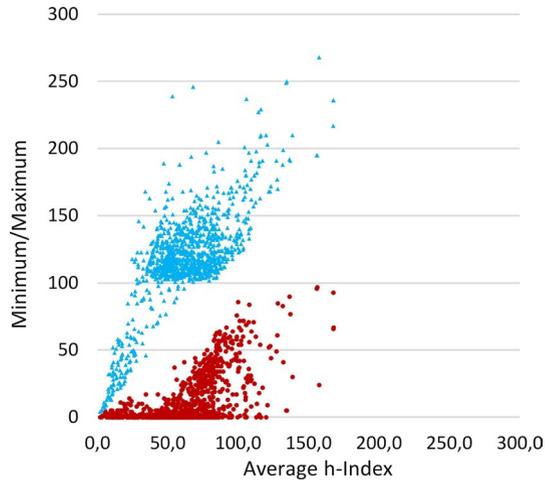


Fig. 4 Box plots of the  $h$ -indices for the names in  $D_{TUKL}$

The full scope of the individual discrepancies of  $h$ -index values (RQ2) becomes apparent when considering the individual maximum and minimum  $h$ -index value that can be found on any of the five platforms. Naturally this discrepancy is somehow related to the actual  $h$ -index, which however can only be calculated with great effort. Therefore we have chosen the average  $h$ -index of a person on all five platforms as  $x$ -axis and plotted the minimum and maximum against it (see Fig. 5). The results clearly show that the discrepancies between the  $h$ -indices on the different platforms are enormous even for scientists with

**Fig. 5** Average  $h$ -index against the respective maximum (blue) and minimum (red)  $h$ -index on all five platforms for every person in the datasets. (Color figure online)



a small  $h$ -index. The potential harm of consulting the  $h$ -index on the wrong platform is considerably high, since an  $h$ -index that is too small can have a negative impact on a scientist's career, whereas an  $h$ -index that is too large can lead to unfair competition and thus to an advantage for scientists who deserve it less than their competitors. In the area of individual observation, further studies in quantitative terms do not exist yet. Thus, our results harden the impression that the individual database issues have a higher negative impact than assumed by studies examining small datasets like Bar-Ilan (2008).

Since a comparison with the approximately correct real  $h$ -indices disambiguates the magnitude of differences even further, we elaborate them for the scientists in dataset  $D_{\text{TUCL}}$  in study III.

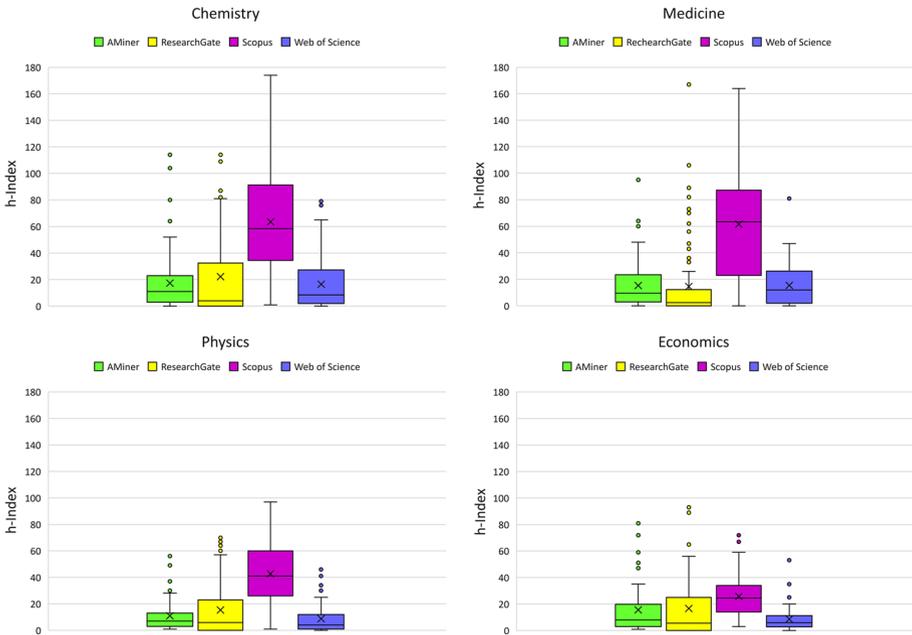
## Study II

### Description

With the help of data set  $D_{\text{nobel}}$ , an analysis of the importance of the various scientific disciplines of the Nobel laureates for their  $h$ -index (RQ3) was carried out on several platforms. Since only 89 names are too few for a meaningful analysis, we investigated which circumstances have led to such a reduction from previously 632 names.

### Method

*Step 1* Reevaluation of Study I: To investigate why only 89 names remained, we examined the refined results from Study I with regard to how many results could be found on each platform. It turns out that on Google Scholar noticeably less names could be found than on the others, therefore Google Scholar is excluded from this study. 289 names remain for which results on the other four platforms could be found. *Step 2* Division into scientific disciplines: The results found are split into the four Nobel Prize categories considered. Of the 289 remaining laureates 78 have a Nobel Prize in chemistry, 90 in medicine, 71 in physics



**Fig. 6** Box plots of the *h*-indices for the Nobel laureates in chemistry, medicine, physics and economics

and 50 in economics. People who have won a Nobel Prize in two different categories<sup>2</sup> are listed once per category.

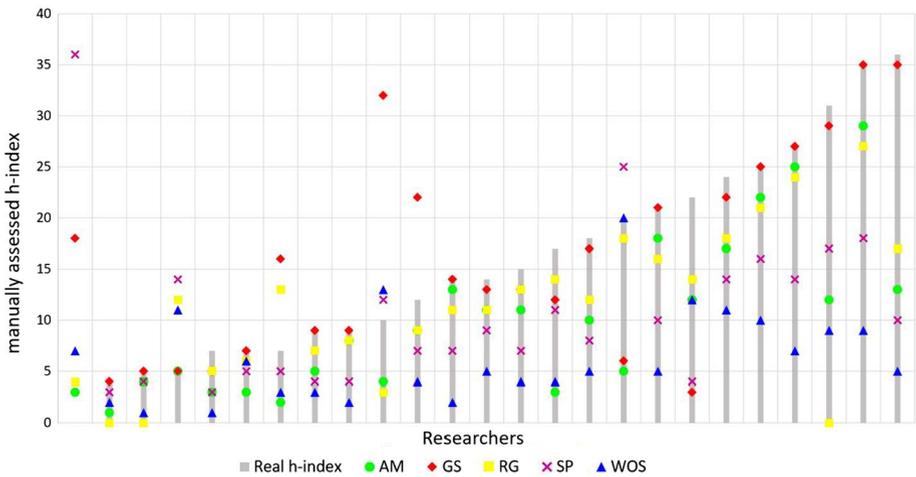
**Results**

The results for the fields of chemistry and medicine are very similar to the overall results for  $D_{\text{nobel}}$ , while those for physics and economics are significantly lower (see Figure 6).

**Discussion**

The finding corresponds to the results of comparable analyses (like for example Iglesias and Pecharrómán (2007)), but here, too, the significance of the data must be critically assessed, since the data set is limited to the names of Nobel laureates who do not necessarily have an average publication behaviour. Additionally, winning a Nobel Prize necessarily results in an increase of popularity of the scientist, which could trigger atypical citation rates for the respective scientific discipline.

<sup>2</sup> For example Marie Curie won in 1903 together with Henri Becquerel in physics and in 1911 in chemistry.



**Fig. 7** Plotted results of the manually assessed  $h$ -index and the corresponding  $h$ -indices on the five platforms for the 25 names from  $D_{TUKL}$

## Discussion

### Study III

#### Method

Of great interest is the question of how well the platforms approximate the truth, i.e., the real  $h$ -index of a person (RQ4). Of course, assessing this real  $h$ -index is a problem on its own: The person needs to know all her publications and all references made to them from all valid, scientific documents. To tackle this question, we collaborated with 27 out of the 32 people in  $D_{TUKL}$  to evaluate their  $h$ -index by first using the Publish or Perish tool with subsequent assessment by hand. The Publish or Perish tool by Anne Harzing (2007) is based on Google scholar which tends to assign too many papers to persons rather than too few. By going through the list of publications assigned by the tool to the scientists, we deleted those that clearly did not belong to the scientists. The process led to applicable results for 25 scientists.<sup>3</sup> It is still possible that publications or citations were missing. Thus, —even with an assessment by hand—the  $h$ -index might still not be the real one, however, as it was assessed together with the scientists we assume that the deviation should not exceed 2–5 points.

<sup>3</sup> The Publish or Perish tool allows only 1000 results. For two names this limit has been reached before finding one correctly assigned publication.

**Table 7** Results of the manually assessed  $h$ -index and the corresponding  $h$ -indices on the five platforms for the 25 names from  $D_{TUKL}$  and the sample standard deviation for each platform as described in Eq. 1

	Manually assessed $h$ -index	AMiner	Google Scholar	ResearchGate	Scopus	Web of Science
Person 1	0	3	18	4	36	7
Person 2	4	1	4	0	3	2
Person 3	5	4	5	0	4	1
Person 4	5	5	5	12	14	11
Person 5	7	3	5	5	3	1
Person 6	7	3	7	6	5	6
Person 7	7	2	16	13	5	3
Person 8	9	5	9	7	4	3
Person 9	9	8	9	8	4	2
Person 10	10	4	32	3	12	13
Person 11	12	9	22	9	7	4
Person 12	14	13	14	11	7	2
Person 13	14	11	13	11	9	5
Person 14	15	11	13	13	7	4
Person 15	17	3	12	14	11	4
Person 16	18	10	17	12	8	5
Person 17	20	5	6	18	25	20
Person 18	21	18	21	16	10	5
Person 19	22	12	3	14	4	12
Person 20	24	17	22	18	14	11
Person 21	25	22	25	21	16	10
Person 22	27	25	27	24	14	7
Person 23	31	12	29	0	17	9
Person 24	35	29	35	27	18	9
Person 25	36	13	35	17	10	5
Sample standard deviation ( $d_p$ )		8.5294	8.1394	8.7011	12.4783	13.3401

## Results

The results are sorted by the manually assessed  $h$ -index and displayed in Fig. 7. Table 7 lists the exact values and shows the sample standard deviation  $d_p$  for each platform  $p$  computed by

$$d_p = \sqrt{\frac{\sum_{i=1}^N (h_{p,s_i} - \bar{h}_{s_i})^2}{N - 1}} \tag{1}$$

where  $N$  represents the number of scientists,  $h_{p,s_i}$  represents the  $h$ -index of scientist  $s_i$  on platform  $p$  and  $\bar{h}_{s_i}$  represents the respective manually assessed  $h$ -index. To investigate the extent of differences even further, we also compared the manually assessed  $h$ -index for each scientist with the respective minimum and maximum value among all platforms and the sample standard deviation  $d_s$  for each scientist  $s$  computed by

**Table 8** Manually assessed  $h$ -index with the respective minimum and maximum value among all platforms and the sample standard deviation for each of the 25 names from  $D_{\text{TUKL}}$  as described in Eq. 2

	Manually assessed $h$ -index	Minimum	Maximum	Sample standard deviation ( $d_s$ )
Person 1	0	3	36	20.5791
Person 2	4	0	4	2.7386
Person 3	5	0	5	3.2787
Person 4	5	5	14	6.4420
Person 5	7	1	5	4.3589
Person 6	7	3	7	2.3452
Person 7	7	2	16	6.3640
Person 8	9	3	9	4.5000
Person 9	9	2	9	4.3589
Person 10	10	3	32	12.0623
Person 11	12	4	22	7.1937
Person 12	14	2	14	7.1239
Person 13	14	5	13	5.5902
Person 14	15	4	13	7.2284
Person 15	17	3	14	10.4283
Person 16	18	5	17	9.6177
Person 17	20	5	25	10.6066
Person 18	21	5	21	10.1366
Person 19	22	3	14	15.4029
Person 20	24	11	22	9.4604
Person 21	25	10	25	9.0967
Person 22	27	7	27	12.0623
Person 23	31	0	29	22.3942
Person 24	35	9	35	16.3172
Person 25	36	5	35	25.1396

$$d_s = \sqrt{\frac{\sum_{i=1}^N (h_{s,p_i} - \bar{h}_s)^2}{N - 1}}, \quad (2)$$

where  $N$  represents the number of platforms,  $h_{s,p_i}$  represents the  $h$ -index of scientist  $s$  on platform  $p_i$  and  $\bar{h}_s$  represents the respective manually assessed  $h$ -index.

## Discussion

As it can be seen, all platforms except for GS yield a too low  $h$ -index value in most cases, while GS overestimates it (see Fig. 7). This observation aligns with the fact that GS provides the most comprehensive database (Jordan 2014; Google Scholar 2017), including multiplied entries and unscientific publications. The frequently correct value estimations might be an indication for accurately maintained profiles on the platform, which is a common thing to do as a university lecturer in Germany. On every platform at least one of the researchers receives a too high  $h$ -index value and another a too low value compared to their manually assessed  $h$ -index values (see Table 7). Thus, it cannot generally be said that one

platform always delivers higher or lower results than expected. Additionally, there is at least one extreme outlier on each platform (GS and SP positive, AM, RG and WOS negative) for this dataset (see Table 8) making it inevitable to properly inspect the results of a name to ensure a minimal deviation.

## Threats to validity

To be able to assess the quality of the results and the limits of their interpretation, we follow a schematic guideline by Drost (2011). This guideline mentions various threats to validity that can be checked in order to gain a better insight into the interpretability of empirical results.

**Internal validity:** Obviously, many names have not been found on all five platforms, which were therefore removed from the study. This applies especially to the names from  $D_{\text{Nobel}}$ , which might lead to biased results. Thus, we have based our findings on the aggregated results of all datasets. Additionally, taking  $h$ -indices computed by hand only from  $D_{\text{TUKL}}$  as reference might be biased to a certain extent, which is why those can explicitly be taken solely as first evidence.

**Construct validity:** Another issue is the way information has been extracted from the platforms. The self-programmed crawler is designed to enter the names and extract the results based on dynamically rendered HTML-structures. Furthermore, for the first study, we always took the first result on that platform with that name while some of them give a selection menu of all persons with the same name. For an automatic tool, it is not possible to choose the most likely correct name.

**External validity:** Our conclusions are based on 1052 names from a variety of four different subgroups, which is why the generalizability to all scientists of the findings might be debatable. Since the results between each subgroup do not vary much, this problem should not be an issue here.

Overall, the list of possible biases is small and most of the items are inevitable for an automated study.

## General discussion

Our results raise the concern, whether promotion committees at universities should use any of the tested platforms to obtain the  $h$ -index of a candidate.

In direct comparison between the platforms, study I shows that the  $h$ -index values yield massive differences (RQ1). Within all datasets, major discrepancies between the respective average  $h$ -indices as well as the interquartile ranges can be observed. These large-scale results strongly substantiate the findings of the small-scale studies mentioned in the introduction and broaden the results of the more specific, disciplinary studies (e.g. Schreiber 2007; Engqvist and Frommen 2008).

The quantitative approach in the second part of study I minimizes the potential impact of random fluctuations and leaves overwhelming evidence that it is impossible to rely on the results of the  $h$ -index provided by only one platform. It validates the previously observed alarming state of the art and indicates a change for the worse.

Study II showed that the ground truth  $h$ -index of 25 scientists (manually assessed) deviated strongly from most of their various online scores to an extent that is not tolerable.

Independent of the strong differences between the platforms, a society that uses such platforms must consider the error tolerance inherent in such automatic performance evaluation processes.

While working on this article we came across three major ethical problems which are caused by the current use of the *h*-index computed by various online platforms:

First of all, there is a significant problem in defining the input. Most platforms do not give clear information about their definition of a 'scientific article'. It is, however, obvious that the different platforms work on very different data bases of articles, based on their definition of that concept. Secondly, a lot of profiles of scientists on AMiner, Google Scholar, ResearchFate and Scopus are created without their consent or even knowledge. Thus, an individual usually does not know how many platforms collect information about himself or herself. If these are used in a job application, for example, an incomplete profile could lead to a rejection without the actual scientific output justifying it. So if a person does not even know that a platform has computed such important metrics, they should at least be informed about it, as well as given the opportunity to correct errors and add data to their profile. Unfortunately, this possibility is often only accessible via a complicated route.

Last but not least, the automated assignment of scientific articles to the right scientists poses a big challenge to a computer. This problem is known as the entity recognition problem, i.e. the attempt to identify the correct entity from properties, in this example from a paper author to the correct individual (O'Neil 2016).

This ties in with a broad ethical discussion about the use of algorithms in various social processes (Lischka and Klingel 2017) and the question when and how algorithms must be made accountable for society (Diakopoulos 2014). As an overview of the responsibilities and sources of error when dealing with algorithms, we refer to Zweig (2018).

Finally, we conclude that as long as the variation between the most commonly used systems to automatically evaluate scientific performance and the real values is as high as measured for this article the systems are neither fair nor neutral and should not be used to assess academic job candidates.

**Acknowledgements** Open Access funding provided by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- AMiner (2017a) AMiner author search facility. Retrieved on May 15, 2017 from <https://aminer.org/>.
- AMiner (2017b) AMiner researcher with highest *h*-index. Retrieved on May 7, 2017 from [https://aminer.org/ranks/experts/h\\_index](https://aminer.org/ranks/experts/h_index).
- Ball, P. (2007). Achievement index climbs the ranks. *Nature*, 448, 737.
- Bar-Ilan, J. (2008). Which *h*-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257–271.
- Batista, P. D., Campitelli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179–189.

- Diakopoulos, N. (2014). *Algorithmic-accountability: On the investigation of black boxes*. Tow Center for Digital Journalism. <https://doi.org/10.7916/D8ZK5TW2>.
- Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and perspectives*, 38(1), 105.
- Engqvist, L., & Frommen, J. G. (2008). The *h*-index and self-citations. *Trends in Ecology & Evolution*, 23(5), 250–252.
- Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB Journal*, 22(2), 338–342. <https://doi.org/10.1096/fj.07-9492lsf>.
- Gapski, H. (2015). Big data und medienbildung. *Zwischen Kontrollverlust, Selbstverteidigung und Souveränität in der digitalen Welt*. Düsseldorf-München.
- Google Scholar. (2017). Helping menu on Google Scholar. Retrieved on May 6, 2017 from <https://scholar.google.de/intl/de/scholar/help.html#coverage>.
- Harzing, A. W., et al. (2007). *Publish or perish*. <https://harzing.com/resources/publish-or-perish>.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569.
- Iglesias, J. E., & Pecharromán, C. (2007). Scaling the *h*-index for different scientific isi fields. *Scientometrics*, 73(3), 303–320.
- Jordan, K. (2014). Academics' awareness, perceptions and uses of social networking sites: Analysis of a social networking sites survey dataset. <https://doi.org/10.2139/ssrn.2507318>.
- Lischka, K., & Klingel, A. (2017). *Wenn Maschinen Menschen bewerten*. Bertelsmann-Stiftung: Arbeitspapier. <https://doi.org/10.11586/2017025>.
- MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5), 342.
- Nature. (1965). Science citation index. *Nature*, 205(4970), 450–450. <https://doi.org/10.1038/205450d0>.
- Nature. (2017). Credit where credit is due. *Nature*. <https://doi.org/10.1038/462825a>.
- Nobel. (2017). List of all Nobel laureates. Retrieved May 7, 2017 from [https://www.nobelprize.org/nobel\\_prizes/](https://www.nobelprize.org/nobel_prizes/).
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing Group.
- Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., et al. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, 111(43), 15316–15321.
- Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493(7431), 159.
- Razak, R. A., & Fahrurazi, F. R. (2011). Agile testing with selenium. In: *2011 5th Malaysian conference in software engineering (MySEC)* (pp. 217–219). IEEE.
- ResearchGate. (2017). ResearchGate author search facility. Retrieved on May 07, 2017 from <https://www.researchgate.net/>.
- Reuters. T. (2008). Web of science: The definitive resource for global research. *linea*. *Disponibile en: scientific.thomsonreuters.com*.
- Saleem, T. (2011). The Hirsch index-a play on numbers or a true appraisal of academic output? *International Archives of Medicine*, 4(1), 25.
- Sattelberger, T. (2015). Abhängiger oder souveräner Unternehmensbürger – der Mensch in der Aera der Digitalisierung. In T. Sattelberger, I. Welpel, & A. Boes (Eds.), *Das demokratische Unternehmen* (pp. 33–53). Freiburg, München: Neue Arbeits- und Führungskulturen im Zeitalter digitaler Wirtschaft.
- Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78(3), 30002.
- Scopus. (2017). Scopus author search facility. Retrieved May 07, 2017 from <https://www.scopus.com/free/lookup/form/author.uri>.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 990–998). ACM.
- Web of Science. (2017). Web Of Science author search facility. Retrieved May 7, 2017 from [https://apps.whoofknowledge.com/WOS\\_GeneralSearch\\_input.do?product=WOS&search\\_mode=GeneralSearch](https://apps.whoofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch).
- Webometrics. (2017). Webometrics Ranking of World Universities. Retrieved on May 30, 2017 from [www.webometrics.info/en/About\\_Us](http://www.webometrics.info/en/About_Us).
- Zweig, K. A. (2018). Wie Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. *Bertelsmann-Stiftung Arbeitspapier*. <https://doi.org/10.11586/2018006>.