Scientometrics (2020) 125:2505–2522
https://doi.org/10.1007/s11192-020-03691-3

Should Google Scholar be used for benchmarking against the professoriate in education?

Margaret K. Merga[1] · **Sayidi Mat Roni**[2] · **Shannon Mason**[3]

Received: 4 May 2020 / Published online: 14 September 2020

Abstract
In the neoliberal environment of contemporary academia, an individual's research rankings and outputs can shape their career security and progression. When applying for ongoing employment and promotional opportunities, academics may benchmark their performance against that of superior colleagues to demonstrate their performance in relation to their discipline. The H-index and citation rates are commonly used to quantify the value of an academic's work, and they can be used comparatively for benchmarking purposes. The focus of this paper is to critically consider if Google Scholar be used for benchmarking against the professoriate in education, by weighting up issues of data reliability and participation. The Google Scholar profiles of full professors at top ranked universities in Australia, the United Kingdom and the United States of America are analysed to explore how widespread Google Scholar use is in the education professoriate. Quartiles of impact are established in relation to H-index, with exploration of how gender is distributed across these quartiles. Limitations of using Google Scholar data are highlighted through a taxonomy of quality confounders, and the utility of Google Scholar as a legitimate tool for benchmarking against the professoriate in education is strongly challenged. As metrics continue to rise in their importance for academics' job security and promotional prospects, reliance on metrics of dubious quality and uneven participation must be questioned.

Keywords H-Index · Benchmarking · Education · Google Scholar · Gender

Introduction

In contemporary academia, an individual's research rankings and outputs shape their job security and progression in a highly competitive environment (Osterloh and Frey 2015). Considering the increasing instability of employment in academia, characterised by scarcity of ongoing roles, it can be contended that "those of us fortunate enough to hold

✉ Margaret K. Merga
m.merga@ecu.edu.au

[1] School of Education, Edith Cowan University, Perth, Australia

[2] School of Business and Law, Edith Cowan University, Perth, Australia

[3] Faculty of Education, Nagasaki University, Nagasaki, Japan

tenured positions at financially stable universities may be the last faculty to enjoy such comparative privilege" (Guthrie et al. 2015, p. 3). Full professorship is the ultimate goal for many academics in the discipline of education, though to secure initial employment may be a more immediate and pragmatic concern. Before the goal of full professorship can be attained, academics must secure ongoing, secure employment through tenure, and then progress upward through internal promotion or mobility between institutions. The case put forth to surpass all of these hurdles is typically an argument for merit that may to some extent be reliant on benchmarking against the performance of other academics in a discipline. Benchmarking is felt to be particularly valuable where the deciding committee includes individuals from outside the applicant's discipline, as it can be difficult to determine the degree of research productivity that is considered sufficient for advancement in other disciplines (Glover et al. 2012). Where applicants benchmark by comparing themselves against more senior colleagues within their discipline, they can show that they are meeting or exceeding disciplinary norms in research output production. However, while it is a common feature of such cases, little attention is given to how this benchmarking should be performed, and the implications of the activity.

The H-index

While there is no known universally-accepted manner in which to present this case of merit, metrics such as the H-index and citation rates are commonly used to quantify the value of an academic's work, and they can be used comparatively for benchmarking purposes. While citation rate is a count of how many times a particular piece of work has been cited in other work, the H-index is calculated as follows: "A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np h) papers have no more than h citations each" (Hirsch 2005, 16,569), meaning that "a researcher has an H-index of 20 if he or she has published 20 articles receiving at least 20 citations each" (Barnes 2014, p. 456).

We also note that the citation count that underpins the H-index is itself a metric that can be shaped by various, often unpredictable factors (Hartley 2019). In some cases, dubious citation farming practices can lead to the inflation of citation counts, though a high rate of self-citation is not always related to unethical research behaviour (e.g. Van Noorden and Chawla 2019). Self-citation patterns are also influenced by gender, with the disparity between men's and women's citation rates growing over time. King et al. (2017) observe that "in the past two decades, we find that men cite themselves 70 percent more, compared with 57 percent more across the more than two centuries of our full data set" (p. 4).

The H-index rewards a certain kind of success in academia. As noted by Costas and Franssen (2018), "the most important challenge of the h-index is that, like essentially any other single indicator, it introduces a particular notion of scientific performance as ideal" (p. 1128). Academics with a smaller number of very highly cited papers will have a smaller H-index than those with a large number of moderately cited papers, leading to claims that the H-index rewards mediocrity (Barnes 2014). However, its popularity may align with its simplicity. Hirsch (2005), the creator of the H-index, contends that it is "an easily computable index, *h*, which gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions", and that it "may provide a useful yardstick with which to compare, in an unbiased way, different individuals competing for the same resource when an important evaluation criterion is scientific achievement" (p. 16,572).

Benchmarking research outputs of educational researchers is a controversial endeavour. Rankings drawn from research outputs and citations influence research governance and are used to assess the performance of academics, departments, and institutions (Osterloh and Frey 2015). The widespread use of metrics such as the H-index is reflective of the increasingly dominant neoliberal forces in higher education and research. The H-index "has become reified; it has taken on a life of its own; a number that has become a rhetorical device with which the neoliberal academy has come to enact 'academic value'" (Burrows 2012, p. 361). In this metric-driven environment, the autonomy of academic researchers is called into question.

> Are academics free to pursue their own research interests, in this metric driven and controlled environment? Or, as national priorities for research are identified and embedded within university systems and academics are increasingly measured and assessed by defined research outputs, would they rather conduct safe rather than speculative or contentious research… (Martin-Sardesai et al. 2017, p. 380)

To use such metrics is to be complicit in the commodification of academic work, with academics themselves becoming commodities, and producers of research outputs and rankings (Martin-Sardesai et al. 2017). The reliability of the H-index as a measure of performance has also been strongly contested in recent times (Barnes 2014), and it may reward unethical practices in academia. For example, it is also possible to manipulate the H-index, such as by using "collusive citation via citation cartels, wherein scholars create informal agreements to surgically monitor and cite member articles" (Haley 2017, p. 87).

The H-index obscures issues of inequity that often lie at the root of difference in individual performance. Capacity for individual attainment in metrics is not even, as research suggests that a variety of extrinsic factors can shape academics' capacity to publish. While Hirsch (2005) suggests that the H-index is "unbiased" (p. 16,572), production of research outputs is reflective of privilege in academia, with factors such as gender, age, ethnicity and nation of employment shaping an individual's attainment. For example, a report on gender in relation to research participation and career progression and perceptions by Elsevier (2020) found that while there has been some progress toward achievement of gender parity in many fields, "the ratio of women to men as authors decreases over time, contributing to men publishing more, having greater impact as well as exposure to international career advancement" (p. 6).

Tools for H-index calculation

A number of different tools can be used to calculate citation rate and H-index, though each has noteworthy limitations in relation to scope of inclusions and access (Mahé 2017). Choice of citation tool to calculate H-index of researchers is important, as there are disciplinary differences in the breadth of database coverage, which advantage some disciplines (Bar-Ilan 2008). Commonly used tools include, but are not limited to Web of Science (WoS), Scopus, ResearchGate and Google Scholar (GS), and previous research has found dramatic differences in H-indexes for a single individual across these four tools (e.g. Teixeira da Silva and Dobránszki 2018). Whereas Scopus provides citation data solely for outputs indexed in it, and both Scopus and WoS are selective databases limited to journals and conference proceedings that fit their quality criteria (Bar-Illan 2018), ResearchGate and GS are broader in their inclusions, and allow for greater researcher control over inputs.

Not all scholarly text types are equally privileged across these tools. Scopus and WoS do not generally cover books and book chapters, and while "Scopus includes some book chapters", "the emphasis in both of these databases are on journal publications" (Bar-Illan 2018, p. 1116). Bar-Illan (2018) argues that this privileging of journal articles is important to ensure quality is valued, however this stance is reflective of disciplinary norms that may be alien to academics in the discipline of education, where peer-reviewed academic research books are viewed as quality research outputs. While scholars in the humanities and social sciences increasingly publish journal articles, books and book chapters are traditional avenues of publication for these individuals, whereas in contrast, "the major means for disseminating research among natural scientists and mathematicians has been journal articles" (Sabharwal 2013, p. 142). As such, it cannot be contended that Scopus and WoS are reasonable metrics to draw from when seeking to evaluate the performance of the education professoriate, who fall within the humanities and social sciences. In addition, full professors have commonly been publishing for many years, and "late career stage faculty members are disadvantaged as a result of changing productivity output—they produce more books over time" (p. 158). It should also be noted that woman may publish more books and men more articles, and therefore weighting journal articles as a superior output can have consequences for gender parity in academia by making it harder for women to compete for promotional opportunities (Millar and Barker 2020).

Google Scholar

GS has several features that can be attractive to educational researchers. Researchers create an "editable, verified (using an institutional email) profile including their personal details, a list of their papers, and citations to those papers" (Bar-Ilan et al. 2012, p. 5). Ortega (2015) notes that the appeal of GS lies in that

> it makes possible the definition of specific research units, mainly researchers, which are able to be compared with others inside the same institution or research interest. In addition, the comprehensive coverage of research materials in GS favours that these pages offer a wide view of the research production and impact. And finally, the fact that these profiles are publicly available, it helps that an author can be appreciated for a broader range of academic activities. (p. 2)

However, he also acknowledged the issues inherent in affording researcher some control over their profiles, which are created and made publicly accessible by researchers. The population of GS may similar to academic social networking sites, and therefore it may yield unbalanced samples in relation to discipline, institution, and nation, both currently and over time.

The design of GS lends itself to manipulation, which is both an advantage and a disadvantage. As explained by Van Bevern et al. (2016), despite the automatic computation of profiles, owners can alter their H-index though merging articles, and while the purpose of this is to combine alternate versions of the same article, this function can be abused by academics seeking to artificially inflate their H-index.

> For example, a researcher may want to merge a journal version and a version on arXiv.org, which are found as two different articles by Google's web crawlers. This may decrease a researcher's H-index if both articles counted towards it before merging, or increase the H-index since the merged article may have more citations than each of the

> individual articles. Since the Google Scholar interface permits to merge arbitrary pairs of articles, this leaves the H-index of Google Scholar profiles vulnerable to manipulation by insincere authors. (p. 20).

However, little is known about the extent to which this functionality is exploited for this purpose. Researchers have also demonstrated the relative ease with which it is possible to index fake papers in GS and manipulate citation data (Delgado López-Cózar et al. 2014). Considering the value that is given to these metrics, and their potential to play a key role in securing ongoing employment or promotional opportunities, the possibilities of abuse of such affordances warrants further examination.

GS was selected as the focus of this paper on benchmarking as it is more accessible than WoS and Scopus. In comparison with Scopus and WoS, "GS offers advantages in cost (free) and breadth of coverage (the entire Internet), at the expense of the inclusion of fringe material" (Albion 2012, p. 225). In addition, differences in citation rates between disciplines are significantly less pronounced in GS than WoS and Scopus (Harzing and Alakangas 2016), rendering it a preferable tool for education academics. However, this does not imply that GS is an optimal tool that can be employed for the purposes of benchmarking in education.

Using Google Scholar for benchmarking against the professoriate

To benchmark against the professoriate in education using metrics from GS, a number of conditions are desirable. If quartiles or levels of attainment are to be developed for benchmarking purposes, such as those developed by Hirsch (2005) himself for the discipline of physics, or Albion (2012) for the discipline of education, the vast majority of the education professoriate would need to hold active profiles on GS that are carefully curated, to ensure that such levels are not skewed through low participation. This previous research by Albion (2012) found that the GS derived H-indexes of a sample of full professors in the Australian professoriate could be grouped into Marginal (H-index of 6), Typical (H-index of 9) and Superior (H-index of 13) categories, however no recent research that compared the performance of full professors in education across nations could be found. It is also crucial that there be greater understanding of how factors such as gender are related to attainment of top quartile status.

Little is known about the percentage of the education professoriate who maintain GS profiles. High participation cannot be assumed, as extant research suggests that it could be less frequently used to host a profile than ResearchGate and LinkedIn (Greifeneder et al. 2018). Earlier research based on a small sample of researchers presenting at a science, technology and innovation conference found that only 23% used GS profiles, though the authors suggest that this low rate could be reflective of the relative newness of the tool at the time; GS launched its profile service 'Google Scholar Citations' on November 16, 2011 (Bar-Ilan et al. 2012). Different disciplines may not equally prioritise the adoption of GS (Mas-Bleda et al. 2014; Ortega 2015). As such, it is a crucial starting point that level of participation of the education professoriate be examined.

Research Questions

This paper draws on this publicly accessible GS data from full professors working in Australia (AU), the United Kingdom (UK) and the United States of America (USA). The research questions investigated in this paper are:

1. How widespread is GS use in the education professoriate?
2. Does university ranking influence GS use?
3. Is the H-index value purely related to time?
4. What quartiles of impact can be established in relation to H-index?
5. How is gender distributed across these quartiles?
6. Should GS data be used for benchmarking against the professoriate in education?

As such, our paper critically explores if GS should be used for benchmarking against the professoriate in education, by weighting up issues of data reliability and participation. It also explores the extent to which H-index can be related to length of time publishing, quartiles of impact that can be used to classify the performance of professors, and the relationship between gender and research performance in GS H-index.

Method

Identifying universities

The discipline of education was selected as the focus as Merga and Mason work in this discipline, and all three authors conduct research in this discipline. Therefore, the authors are familiar with disciplinary norms in education. In addition, it is intended that education serve as a case study to provide insights that could then be tested in other disciplines. There were also pragmatic purposes behind the paper, with Merga expected to provide benchmarking data to support her 2020 application for promotion, and Merga required to deliver in-house training for her colleagues in the School of Education around benchmarking for promotion and competitive grant applications (as delivered in Merga 2020). This meant that the data were of immediate practical use. With the most recent benchmarking data published in 2012 by Albion, there was a need for current research in this space, as "attempts to measure impact are not particularly useful without discipline-specific performance benchmarks" (Benckendorff and Shu 2019, p. 184).

We focus on full professors in the discipline of education as this is the ultimate goal for many academics in the discipline, and we felt that it would be useful to illustrate the commonalities and differences in this highest echelon and at highest tier institutions, particularly in relation to the quartiles of impact we explore herein. We also acknowledge that the path to full professorship differs between the nations we explore, and therefore full professors in AU, the UK and the USA may not be considered to be exactly equivalent (see Table 1 in Benckendorff and Shu 2019, p. 187).

To collect a corpus of data on the GS performance of full professors in AU, the UK and the USA, Quacquarelli Symonds (QS) rankings of universities were used to determine the top 20 universities in each of these three nations. These highest-ranked universities were selected to secure a sample representative of optimal attainment in the professorship. Japan was initially included as a non-Western nation for comparison, but unfortunately very few academics in high-ranked Japanese institutions could be found with GS profiles, and therefore Japan was ultimately excluded.

After identifying the top 20 universities in each nation, course offerings at these universities were investigated, and universities without a School/Department/Faculty of Education were excluded. An online list of the staff (known in the USA as faculty) was then sourced, and universities without clearly identifiable full professors in their School/

Department/Faculty of Education were excluded. Lack of clearly identifiable full professors could be due to the university not having full professors employed in their School/Department/Faculty of Education, or because they were not (clearly) listed online. Finally, universities with full professors, none of whom had GS profiles, were also excluded. The frequency of these exclusions is detailed in Table 1 below. Where an exclusion was applied, the next university on the QS rankings was added, to arrive at the top 20 universities for each nation that met the criteria.

Collecting data on the professoriate

The GS profiles of full professors at 20 institutions across three nations were individually searched. Due to the prevalence of common names in academia, only profiles that had the correct affiliation were included. To ensure only profiles that the professors themselves were aware of were used, only GS profiles which had a verified email associated were included. When creating the list, all full professors were included, with the exception of honorary and clinical professors, adjunct only staff, and emeritus professors. In some unusual cases, a full professor was listed amongst education staff, but a review of their publications showed that they did very little or no work in any area relating to the broad discipline of education, in which case they were excluded. This substantial body of professors was needed to ensure the significance of the findings within a common field.

Data on the following were manually collated: university; name; date of first publication; date of last publication; span of years publishing; total number of citations; citations since 2015; H-index; and, H-index since 2015. An assumed gender was assigned for each individual. Gender was appointed based on dated dichotomous hegemonic norms, where gender was determined by gender traditionally ascribed to names. Where the individual had a gender-neutral or unisex name, additional information was sought, predominantly use of gendered pronouns in online institutional biographies.

Though use of Harzing's (2007) Publish or Perish software can be used for analysis of GS data, this tool was not used, and all data were entered manually. This approach was favoured as a range of publication data also needed to be collected, and the authors had sufficient familiarity with errors in GS to value dealing with the data directly in its primary form. To this end, data exactly reflected the online data available through GS at the time of data collection, with one important exception further explored in the limitations component of this paper. In brief, in many cases, the date of first publication was highly implausible. A number of living professors in education who also had recent publications began publishing over 100 years ago according to the GS profiles, clearly indicating an error.

This meant that at data entry phase, a logic test was applied. It was not assumed that early publications in different disciplines were not the work of the professor, as researchers can and do move between disciplines making use of their transferrable skill sets. Early

Table 1 University exclusions

Category of exclusion	AU	UK	USA
No School/Department/Faculty of Education	1	7	6
No identifiable full professors	1	0	0
No full professors with GS profiles	0	0	2
Total excluded	2	7	8

attributed publications on professor's profiles was only excluded where authorship was not possible (such as where date of first publication was in the 1700s), or in the case of out of discipline work that did not match the name. Again, care needed to be taken here to ensure that names that changed through marriage or gender transition did not inadvertently lead to inappropriate exclusions, so features such as changing middle initials or lack of clear authorship on the work itself were also drawn upon to constitute grounds for exclusion in these cases. As a result, while data collection was time-consuming, the span of publications reflected in the data in this paper is more accurate than those presented in the often poorly maintained GS profile pages. This highlights why desk-based rather than machine-based research may still be ideal for conducting analyses of data sets that require such logical interpretation.

Methodological limitations

Discussion of quality issues in relation to GS is not new. Academics are required to review and edit the materials that GS attributes to them in order to ensure the accuracy of their profiles, and they do not always do so. Acknowledged concerns include a substantial volume of duplicate papers included in GS, termed "stray citations", "where minor variations in referencing lead to duplicate records for the same paper" (Harzing and Alakangas 2016, p. 802), which can lead to inflation of H-index and citation count. During this research process, recurrent serious issues that are quality confounders in GS were found, making it desirable to follow up this project with content analysis that more closely addresses these issues. Some recurring examples are presented herein to inform this future research goal. However, profile identities are not disclosed, to avoid identifying and therefore repudiating the individuals whose profiles contained these errors.

The issues raised herein posed challenges when seeking to establish the *range of publication* (earliest to latest) because as aforementioned, there were many issues with the reliability of the earliest publications listed on GS. As such, errors that did not impact on this determination are not considered here (such as stray citations), so this list is not a comprehensive coverage of all quality issues relating to GS, rather those centrally concerned with this facet of our method. They illustrate that inclusions and exclusions of publications was a complex and sensitive issue, which would not arise if the academics themselves carefully curated what appears on their GS profile, which they have power to do. As such, the limitations section of this paper presents a taxonomy of quality confounders.

Implausible date range

The GS profile of one of the most highly cited researchers included a patent from 1897 that had been cited twice. As the academic was still research active, a publication span of more than one hundred years was considered implausible, and 1897 was not the designated starting point for this academic's publication range. Another active researcher produced a publication with recommendations on education technology that was ostensibly published in 1890; both the topic and the date made this unlikely. The oldest early publications attributed to a living, active author that emerged in our process was a 1795 work on magnetic field dependence of ultracold collisions, a 1787 work on the Scots musical museum, and a 1771 publication on Thomas Gage papers. Even earlier were a year 86 work on state test scores in Texas, which apparently was written before Texas was a state, and a very

implausible year 9 work on children as game designers apparently published 2011 years ago.

Multiple dates

Some items made reference to multiple dates. For example, a researcher had a 1917 work that seemed unlikely to belong to her. It seemed to be a misdated work that was actually produced in 1991; the GS attributed year was 1917, but in parentheses on the profile page was the 1991 date. The 1991 date was considered the more reliable of the two, though the source was also checked for confirmation.

No mention of the author

In some cases, publications were attributed to an author that bore no mention of any of the authors names in the author list. Unless the individual changed their first and last names (and middle initials), these attributions were considered likely to be incorrect, leading in some cases to artificial inflation of citation rates and H-indexes. Perhaps the strangest example of this encountered in the data set was an academic who had a 1989 journal article authored by "Impaler Vlad" (commonly known in popular culture as Dracula) attributed to them. Unsurprisingly, this was not the listed name of the academic. It is possible that some academics completely changed all of their names, but it was considered unlikely, so an on-balance decision was made that a completely changed name was grounds for publication exclusion in this study. However, short forms were taken into account. For example, if Amanda Smart became Mandy Smith, this was considered possible, with Mandy a short form for Amanda, and Smith a possible changed/married last name. We were also aware of the numerous issues raised in GSs' indexing of works of authors with compound last names (e.g. as explored in da Silva and Dobránszki 2018a, b).

Gender change

Great care needed to be taken when the first publication was attributed to an individual of a different gender, as gender transition or fluidity may have occurred, so exclusion was not automatic. For example, where the work of Robert was listed in the profile of Rhonda (same last name), this early publication was only excluded if there were other factors that made the publication unlikely, such as a relatively early date (1969) where the academic was known to the researchers as an individual extremely unlikely to be more than fifty years old.

Different field

As per gender change, publishing out of field was not a grounds for exclusion as many academics do this. However, when it was combined with other factors, such as no mention of the author, or an implausible date, it was grounds for exclusion.

Results

How widespread is Google Scholar use in the education professoriate?

As a result of this approach, a sample of $N=788$ professors was attained as per Table 2 below.

Does university ranking influence Google Scholar use?

In order to determine whether the professoriate's presence on GS is related to their university's ranking, four point-biserial correlation tests between the university ranks and the presence of the professors on GS ($N=788$) were run. The tests were bootstrapped with *bias corrected accelerated* (BCa, $N=1000$) confidence intervals. Bootstrap was selected to control for the non-parametric data distribution (Efron 1981), and BCa was used as this method provides more stable estimates (Kelley 2005).

The results suggest that that while presence on GS is statistically related to the university ranking across the whole set $r_{pb}=.123$, $p<.001$, $N=788$, country-wise associations were different in terms of statistical significance, magnitude, and directions of the relationships. In AU for example, there was no evidence to suggest that the university ranking was associated with the presence on GS, $r_{pb}=.103$, $p=.181$, $n=170$. Similarly, UK also did not support the proposition, $r_{pb}=-.043$, $p=.548$, $n=200$.

The only country that showed a significant association between the university ranking and use of GS was US, $r_{pb}=-.139$, $p<.001$, $n=418$.

Is the H-index value purely related to time?

In Table 3 below, the highest and lowest individual scores found within nations are detailed, along with the median.

A Spearman *rho* test was run to investigate if the years of publication is associated with the number of citations (Mat Roni, Merga and Morris 2020). The result showed that there was a moderate rather than strong statistically significant correlation between the range of years publishing and the number of citations received, $r_s=.47$, $p<.001$ (Taylor 1990). The years publishing accounts for 22% of variation in the number of citations.

What quartiles of impact can be established in relation to H-index?

Data were rendered into quartiles to enable levels of achievement or impact as per Hirsch (2005) and Albion (2012). From the sample ($N=397$), data were distributed as

Table 2 Full professors GS profile adoption

Profile status	AU	UK	USA
With GS profile	118	98	181
Without GS profile	52	102	237
Total	170	200	418
% with GS	69.41	49.00	43.30

Table 3 Range in years publishing, citations, H-index, H-index (since 2015)

Profile status	AU	UK	USA	Total, $N=397$
Longest years publishing	55	58	62	62
Shortest years publishing	12	8	7	7
Median years publishing	26.50	27	28	27
Highest number of citations	92,089	52,479	96,007	
Lowest number of citations	152	2	456	3
Median number of citations	2461.50	3035.50	6960.00	4554
Highest H-index	109	84	115	115
Lowest H-index	6	1	9	1
Median H-index	24	28	35	30
Highest H-index (since 2015)	72	59	68	72
Lowest H-index (since 2015)	6	1	3	1
Median H-index (since 2015)	19	21	25	22

per Table 4, making visible the minimum requirement H-index attainment to achieve each quartile. Albion's (2012) earlier nomenclature are drawn on (as described in the introduction). As such, we designate these as Marginal (lowest quartile, Q4), Moderate (second-lowest quartile, Q3), Strong (second-highest quartile, Q2) and Superior (highest quartile, Q1).

How is gender distributed across these quartiles?

Results from a bias corrected accelerated (BCa) bootstrap *t* test indicates that there were statistically significant differences in gender in terms of H-index, and H-index since 2015, with male academics slightly higher than their female counterparts for both indices (see Table 5 below).

Despite Spearman *rho* test suggesting that the years publishing was correlated with H-index ($r_s=.47$, $p<.001$) and H-index since 2015 ($r_s=.30$, $p<.001$), the publishing span was similar between male and female academics, $t(355.37)=-1.68$, $p=.09$. We summarise the results in Table 5.

Table 4 Lower bounded limits for H-index quartiles of full professors in AU, UK and US

Profile status	Lower bound limit
Marginal (Q4)	1
Moderate (Q3)	21
Strong (Q2)	30
Superior (Q1)	43

Table 5 Bootstrap for independent samples test

	Mean		Mean Difference	Bias	Std. Error	Bootstrap		
						Sig. (2-tailed)	BCa 95% Confidence Interval	
	Female	Male					Lower	Upper
H-index	31.30	36.93	−5.634	−.106	1.832	.003	−9.308	−2.235
H-index 2015	23.04	26.36	−3.322	−.083	1.158	.005	−5.566	−1.338
Years publishing	28.37	30.05	−1.688	−.050	.999	.089	−3.477	.142

Levene's test is significant, $p < .05$, hence, equal variances not assumed

Discussion

Findings and limitations discussed raise questions about the legitimacy of using GS data for benchmarking against the professoriate in education. The sample suggests that GS use is not widespread in the education professoriate. As per Table 2, only $n = 397$ (50.38%) of professors in the sample had a GS profile. While this is a strong sample, it cannot be deemed a high participation rate. While an academic does not have to maintain a GS profile to have a GS H-index, we assume that those maintaining a GS profile are more likely to be involved in curating it to ensure that it accurately represents the research impact of the individual within the aforementioned constrains of citation counts and H-Index. However, further research needs to be done to confirm that this is indeed the case, because as explored in this paper, even those with verified GS profiles do not always rigorously curate their profiles.

There were notable differences between nations. As per Table 2, the professoriate in AU were most likely to have a profile on GS, with more than two-thirds having a profile on GS connected to a verified email address. Less than half of the professoriate in the UK and USA had these profiles. This participation rate can perhaps be related to Berlemann and Haucap's (2015) finding of German business scholars disengagement with quantification of their research outputs can to some extent be related to their security of full professorship, however further research across academic levels would be needed to determine if this holds currency in relation to the education professoriate in AU, the UK and the US. As GS use in the education professoriate cannot be viewed as widespread, the pool of participating professors could be skewed toward the more highly research-active and those who will appear in a positive light (e.g. Berlemann and Haucap 2015). That said, these data suggest that this is not exclusively the case, hence the lower bounded limit for H-index sitting at 1 (as per Table 4). With UK and USA participation at less than 50%, and no current evidence that those who take part can be considered uniformly representative of the professoriate, low participation can be seen as grounds for rejecting GS data for benchmarking purposes. It was also found that university ranking was not related to GS use in AU and the UK, though analysis of the US data yielded a significant association between the university ranking and use of GS. This relationship was negative, suggesting that in the US, professors at higher ranked universities do not typically have a stronger presence on GS than their counterparts at lower-ranked universities.

The range of years publishing, number of citations, H-index (overall) and H-index (since 2015) can be particularly valuable for those looking to identify a minimum attainment for

entry into full professorship, to determine if they make the minimum standard. Anecdotally, we have noticed a perception in those occupying the lower levels of academia that reaching professorship is almost unattainable. However, the ranges in h-index that are illustrated in Table 3 can give some comfort, as even at the highest tier universities, some professors hold a h-index as low as 1, similar to single-figure findings from other disciplines reporting full professors with a h-index of 4 (e.g. Benckendorff and Shu, 2019). It was expected that a strong relationship between length of time actively publishing and citation rate would be found; Hirsch (2005) contended that "for a given individual, one expects that *h* should increase approximately linearly with time" (p. 16,569). The relationship was expected to be stronger than the moderate relationship found, suggesting that it is possible for relatively new academics to outperform academics with extensive track records. Expectations of a linear relationship have also been challenged in other research, such as in relation to social scientists in France and Spain, where "results show a non-monotonic pattern for the effect of the authors' career length".

> The H-index first increases with seniority up to a certain level, and then begins to decrease with further increments in career duration. This could be explained by the existence of different patterns for the researchers' productivity. Researchers with lower seniority may be subject to more institutional pressures to publish than those who have already gained some stability in their academic positions. (Dabós et al. 2019, p. 63)

Thus, it is important to note that while length of time publishing likely confers expected benefits for citation rate, it may not fully explain differences in performance.

Levels of achievement or impact were created as per Hirsch (2005) and Albion (2012) with a quartile approach employed. That the bar for marginal attainment (Q4) is set so low (H-index of 1) is a reminder that research performance is not the only criteria determining full professor status, and is perhaps also reflective of changing expectations of the role of full professor over time. Gender was a significant factor at play in quartile attainment. Males were higher achievers despite higher female representation in the sample, with $n=184$ males and $n=213$ females, and men's higher attainment could not be explained due to a longer publishing range. The higher H-index of males is also reflective of previous findings in other fields, such as social work (Carter, Smith and Osteen 2017), and research suggesting that men are more likely than women to be promoted to full professor (e.g. Millar and Barker 2020). While these data suggest that women lag behind their male colleagues in research performance, even though there are more of them in the education professoriate in this sample, more needs to be done to bring the factors underpinning this discrepancy to light.

While there were more women than men in the sample, men outperformed women, and it could not be attributed to greater career length, suggesting that such data and metrics play a role in obscuring inequity in research output performance. Metrics have become part of academic culture and identities, and notions of excellence, both at individual and institutional levels. O'Connor et al. (2020) contend that "excellence is often seen as the dominant logic in decisions about recruitment/progression in higher education. Implicit in this is the idea that excellence is unambiguous, gender neutral and unaffected by context" (p. 196). They also note that "in contexts where public universities are exposed to the conflicting expectations of multiple stakeholders and to increasing pressure as regards accountability, a concept of excellence which is immune from relational, contextual and gender bias is very attractive" (p. 206). Future analysis should also consider race and age as factors that can also act independently or in concert (e.g. Chambers and Freeman 2020).

Differences between the three nations in relation to their academic structures and norms have not been foregrounded in analysis, and they warrant further consideration in future research. In particular, more needs to be learned about the extent to which differences in workloads and resourcing at individual, institutional and national levels both at present and over time influence the unequal output production of the professoriate in this field. Further research should also explore how productivity is influenced by affiliation with different sub-disciplines in the field of education, which may facilitate different levels of attainment and citation rates, as well as other individual factors related to privilege and attainment in education. Further research could also examine the extent to which metrics and benchmarks can enable or constrain mobility between institutions in different research tiers within nations. As noted by Benckendorff and Shu (2019), it is important to avoid "taking citation metrics at face value without considering institutional differences" (p. 188), and these may potentially be even more powerful than differences between nations. In addition, while AU, UK and USA data are reflective of three different national contexts, it can be argued that similar accountability forces are at play. Research in other disciplines has found that "countries with national research assessments (e.g. NZ, Australia, UK) or tenure requirements (e.g. USA, Canada) tend to exhibit higher median h-indices" (p. 189) then those not needing to be as responsive to such accountability requisites. Therefore data from nations less encumbered by these accountability requisites could be juxtaposed for further exploration.

The taxonomy of quality confounders in the limitations records some of the issues with GS. The notion that GS is an imperfect data source is not new; Jasco's (2008) urging that people "think twice before using GS to calculate H-indexes without a massive corroboration of the raw data reported by GS" (p. 451) retains clear currency more than a decade later. However, the taxonomy presented in this paper extends the breadth of articulated concerns. GS profiles that are regularly and carefully curated will have higher accuracy than those that are poorly maintained, and further analysis could seek to quantify the prevalence of the quality confounders raised in our paper and elsewhere.

While GS's coverage and variety of sources may be broader than its counterparts WoS and Scopus, it "collects its data partly by automatically crawling the web without any quality control" (Mas-Bleda et al. 2014, p. 339), and therefore "although most of Google Scholar's results come from publisher websites, its coverage does include low quality ''publications'' such as blogs or magazine articles" (Harzing and Alakangas 2016, p. 802). In contemporary academia, these so-called "low quality" publications do have a different kind of value, as there is increasing impetus for researchers to demonstrate translation and dissemination of their work in such accessible forms (e.g. Merga and Mason 2020), and therefore these contributions, while lacking the rigour of peer-reviewed academic work, may gain value over time, and this value can often also be quantitatively measured through altmetrics. That said, the prevalence of outputs that cannot be deemed scholarly by current standards limits the use of GS, particularly because it is "very time-consuming, to manually clean every academic's record by merging stray citations and removing non-academic publications" (Harzing and Alakangas 2016, p. 802). Questions remain about why many profiles seem to be artificially inflated by erroneous data, though under-reporting is also possible where not all citations are accurately captured and attributed by GS. While this would be expected if participation had been forced, all of the respondents opted into GS when they linked their account to their email in the verification process. It is not known if the issues remain as a result of the time-consuming nature of maintaining an often lengthy GS profile, poor understanding of how to curate the profile relating to gaps in digital literacy, or even

deliberate negligence that can lead to inaccuracies in metrics, and more research needs to be done to explore this.

However, the question of quality in GS analysis must remain secondary to questions around the efficacy of the H-index as a reliable metric, and if there are superior alternatives such metrics must be used. Modifications to the H-index have been proposed in recent years to counter some of the major criticisms it faces (Bornmann et al. 2008). For example, hI annual corrects for career length, addressing the problem of comparing scholars at different career stages (Harzing and Alakangas 2016), while the s-index builds transparency by reporting self-citations (Flatt et al. 2017). The g-index includes a researcher's most cited papers, which may better reflect visibility and lifetime achievement (Egghe 2006). Arandjelovic (2016) offers a modification that gives weighting based on an author's relative contribution to a publication, as determined by their order in the authorship, potentially discouraging the dubious practice of conferring gift, honorary and guest authorship status to an individual who has not adequately contributed to the research (Harvey 2018). Another alternative is the i10-index developed by GS as the number of publications with at least 10 citations (Conner 2011). While each of the indexes has their strengths, the application of any quantitative metric needs to be done with caution, as they do not capture the many attributes that constitute the nebulous concept of research quality, and their limitations must be acknowledged.

Conclusion

As metrics continue to rise in their importance for academics' job security and promotional prospects, reliance on metrics of dubious quality and uneven participation must be questioned. To this end, Barnes (2014) invokes an *emperor's new clothes* analogy, stating that "the costs of relying on a misleading indicator far outweigh the entirely illusory benefits of making more rapid decisions on the basis of false information" (p. 466). Even as this paper mounts a critique, it is acknowledged that the quartiles presented in this paper will almost certainly be used by education academics for benchmarking purposes, as they are required in order to obtain the job security and promotional opportunities that will give academics security into their futures. There is an anguish in this clash between the pragmatic and the ideal. Burrows (2012) contends that "academic value is, essentially, becoming monetized, and as this happens academic values are becoming transformed. This is the source of our discomfort", noting that as academics, "we are fully implicated" in the enactment of this quantification of academic impact (p. 368). As authors of this paper, we are complicit too, especially as we create quartiles of impact for benchmarking, furthering quantification potentialities, though we do so with a critical stance. In the current neoliberal environment, challenging the legitimacy of metrics by making visible their practical and pragmatic limitations is needed, though these limitations need to be understood within a broader critical frame of forces currently dominating higher educational management, policy and discourse.

In addition, we accept that we are mounting a critique against the over-reliance on metrics for research performance evaluation without providing a credible alternative. Indeed, in a number of cases, attempts to implement alternatives in research performance conducted with no or limited reference to bibliometrics have been ineffective, costly and time-consuming, with evaluation highly subjective, and methods for evaluation opaque (Mingers et al. 2017). For example, Britain's Research Excellence Framework (REF) was ineffective,

arduous and expensive, and no better at evaluating research performance than bibliometrics (Sayer 2015). In addition, the REF's peer review processes were far short of international standards, and merely served to "keep academic elites in power" (p. 130). It has been contended that "irrespective of their intrinsic merits or otherwise, metrics would have been an excellent predictor of most universities' performance in RAE 2008 at far lower cost and with far less damaging side effects" (p. 4). As we have explored in our article, using GS for research performance evaluation of the education professoriate is flawed. However, in the absence of viable and fair alternatives that do not disadvantage certain fields and disciplines, it will continue be used. These metrics "do exist and are unlikely to go away, so we should try and ensure that they are as fair, transparent, and harmless as possible" (Mingers et al. 2017, p. 1628). Our data supports the contentions of previous commentators in raising awareness of the importance that researchers curate their profiles to reduce the prevalence of the data quality issues explored in this paper. We hope that this paper can to some extent improve the veracity in representation of GS within its other limitations, and these implications extend beyond the discipline of education. Such metrics will be most useful when viewed as part of a broader academic portfolio that is responsive to sometimes unique disciplinary norms.

References

Albion, P. R. (2012). Benchmarking citation measures among the Australian education professoriate. *The Australian Educational Researcher, 39*(2), 221–235.

Arandjelovic, O. (2016). Fairer citation based metrics. *Publication Research Quarterly, 32,* 163–169.

Bar-Ilan, J. (2008). Which H-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271.

Bar-Ilan, J. (2018). Comments on the Letter to the Editor on "Multiple versions of the h- index: Cautionary use for formal academic purposes" by Jaime A. Teixera da Silva and Judit Dobránszki. *Scientometrics, 115*(2), 1115–1117.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social Web. In *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 98–109). arXiv preprint arXiv:1205.5611.

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*(5), 456–470.

Benckendorff, P., & Shu, M. L. (2019). Research impact benchmarks for tourism, hospitality and events scholars in Australia and New Zealand. *Journal of Hospitality and Tourism Management, 38,* 184–190.

Berlemann, M., & Haucap, J. (2015). Which factors drive the decision to opt out of individual research rankings? An empirical study of academic resistance to change. *Research Policy, 44*(5), 1108–1115.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59*(5), 830–837.

Burrows, R. (2012). Living with the H-index? Metric assemblages in the contemporary academy. *The Sociological Review, 60*(2), 355–372.

Carter, T. E., Smith, T. E., & Osteen, P. J. (2017). Gender comparisons of social work faculty using H-index scores. *Scientometrics, 111*(3), 1547–1557.

Chambers, C. R., & Freeman, S., Jr. (2020). To be young, gifted, and black: The relationship between age and race in earning full professorships. *The Review of Higher Education, 43*(3), 811–836.
Conner, J. (2011). Google Scholar citations open to all. *Google Scholar Blog*. http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html.
Costas, R., & Franssen, T. (2018). Reflections around 'the cautionary use' of the h-index: Response to Teixeira da Silva and Dobránszki. *Scientometrics, 115*(2), 1125–1130.
Da Silva, J. A. T., & Dobránszki, J. (2018a). Multiple versions of the H-index: Cautionary use for formal academic purposes. *Scientometrics, 115*(2), 1107–1113.
Da Silva, J. A. T., & Dobránszki, J. (2018b). Rejoinder to "Multiple versions of the H-index: cautionary use for formal academic purposes". *Scientometrics, 115*(2), 1131–1137.
Dabós, M. P., Gantman, E. R., & Rodríguez, C. J. F. (2019). The prestige of social scientists in Spain and France: An examination of their H-index values using Scopus and Google Scholar. *Minerva, 57*(1), 47–66.
Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology, 65*(3), 446–454.
Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*(3), 589–599.
Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69*(1), 131–152.
Elsevier. (2020). *The researcher journey through a gender lens*. https://www.elsevier.com/research-intelligence/resource-library/gender-report-2020.
Flatt, J. W. (2017). Improving the measurement of scientific success by reporting a self- citation index. *Publications, 5*(3), 1–6.
Glover, S. M., Prawitt, D. F., Summers, S. L., & Wood, D. A. (2012). Publication benchmarking data based on faculty promoted at the top 75 US accounting research institutions. *Issues in Accounting Education, 27*(3), 647–670.
Greifeneder, E., Pontis, S., Blandford, A., Attalla, H., Neal, D., & Schlebbe, K. (2018). Researchers' attitudes towards the use of social networking sites. *Journal of Documentation, 74*(1), 119–136.
Guthrie, J., Parker, L. D., & Dumay, J. (2015). Academic performance, publishing and peer review: Peering into the twilight zone. *Accounting, Auditing & Accountability Journal, 28*(1), 2–13.
Haley, M. R. (2017). On the inauspicious incentives of the scholar-level H-index: an economist's take on collusive and coercive citation. *Applied Economics Letters, 24*(2), 85–89.
Hartley, J. (2019). Some reflections on being cited 10,000 times. *Scientometrics, 118*(1), 375–381.
Harvey, L. A. (2018). Gift, honorary or guest authorship. *Spinal Cord, 56*(2), 91.
Harzing, A.W. (2007). *Publish or Perish*. https://harzing.com/resources/publish-or-perish.
Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.
Jacso, P. T. (2008). The pros and cons of computing the h-index using Google Scholar. *Online Information Review, 32*(3), 437–452.
Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65*(1), 51–69.
King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius, 3,* 1–22.
Mahé, G. (2017). *The indexation of scientific journals and the bibliometry: examples with current tools (Research Note–IRD/HSM Montpellier, France–April 2017)*. https://www.researchgate.net/publication/316191247_The_indexation_of_scientific_journals_and_the_bibliometry_examples_with_current_tools.
Martin-Sardesai, A., Irvine, H., Tooley, S., & Guthrie, J. (2017). Government research evaluations and academic freedom: A UK and Australian comparison. *Higher Education Research & Development, 36*(2), 372–385.
Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. F. (2014). Do highly cited researchers successfully use the social web? *Scientometrics, 101*(1), 337–356.
Mat Roni, S., Merga, M. K., & Morris, J. (2020). *Conducting quantitative research in education*. Berlin: Springer.
Merga, M. K. (2020). *Setting up your academic profiles and benchmarking*. Mt Lawley: Workshop, Edith Cowan University.

Merga, M. K., & Mason, S. (2020). Sharing research with academia and beyond: Insights from early career researchers in Australia and Japan. *Learned Publishing*. https://doi.org/10.1002/leap.1296.
Millar, P. E., & Barker, J. (2020). Gender and academic promotion to full professor in Ontario. *Canadian Journal of Sociology, 45*(1), 47–70.
Mingers, J., O'Hanley, J. R., & Okunola, M. (2017). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics, 113*(3), 1627–1643.
O'Connor, P., López, E. M., O'Hagan, C., Wolffram, A., Aye, M., Chizzola, M. O., et al. (2020). Micro-political practices in higher education: a challenge to excellence as a rationalising myth? *Critical Studies in Education, 61*(2), 195–211.
Ortega, J. L. (2015). How is an academic social site populated? A demographic study of Google Scholar Citations population. *Scientometrics, 104*(1), 1–18.
Osterloh, M., & Frey, B. S. (2015). Ranking games. *Evaluation Review, 39*(1), 102–129.
Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice, 15*(2), 141–163.
Sayer, D. (2015). *Rank hypocrisies: The insult of the REF*. London: Sage.
Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography, 6*(1), 35–39.
Van Bevern, R., Komusiewicz, C., Niedermeier, R., Sorge, M., & Walsh, T. (2016). H-index manipulation by merging articles: Models, theory, and experiments. *Artificial Intelligence, 240,* 19–35.
Van Noorden, R., & Chawla, D. S. (2019). Policing self-citations. *Nature, 572*(7771), 578–579.