Research Evaluation, 26(3), 2017, 169–180
doi: 10.1093/reseval/rvx018
Advance Access Publication Date: 20 May 2017
Article

OXFORD

Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation

Björn Hammarfelt[1,2,*] and Alexander D. Rushforth[2]

[1]University of Borås, Swedish School of Library and Information Science, Allégatan 1, Borås 501 90, Sweden and [2]CWTS, Leiden University, 2333 AL Leiden, The Netherlands

*Corresponding author. Email: bjorn.hammarfelt@hb.se

Abstract

A researcher's number of publications has been a fundamental merit in the competition for academic positions since the late 18th century. Today, the simple counting of publications has been supplemented with a whole range of bibliometric indicators, which supposedly not only measures the volume of research but also its impact. In this study, we investigate how bibliometrics are used for evaluating the impact and quality of publications in two specific settings: biomedicine and economics. Our study exposes the various metrics used in external evaluations of candidates for academic positions at Swedish universities. Moreover, we show how different bibliometric indicators, both explicitly and implicitly, are employed to assess and rank candidates. Our findings contribute to a further understanding of bibliometric indicators as 'judgment devices' that are employed in evaluating individuals and their published works within specific fields. We also show how 'expertise' in using bibliometrics for evaluative purposes is negotiated at the interface between domain knowledge and skills in using indicators. In line with these results, we propose that the use of metrics we report is best described as a form of 'citizen bibliometrics'—an underspecified term which we build upon in the article.

Key words: citizen bibliometrics; economics; biomedicine; judgment devices; Journal Impact Factor; h-index.

1. Introduction

Since the 1970s much of the promise of evaluative bibliometrics (Narin 1976) has been premised on the notion of tempering the subjective and cognitive biases of peer review, so much so that it has often been imagined as an alternative mode of evaluating. In practice however, bibliometrics tends to supplement expert decision-making rather than supplant it (van Raan 1996; Moed 2007). Indeed calls to use (advanced) bibliometrics as part of 'informed peer review' processes have been posited as a means of mitigating the weaknesses of both approaches (Butler 2007). At the same time, there are often assumptions made that simple output indicators like Journal Impact Factor (JIF), h-index, and journal ranking lists are being commonly used in decision-making contexts. Despite such assumptions, to date few have responded to earlier calls by Woolgar (1991) to study actual uses of indicators in peer review and other decision-making arenas. While some attention has been directed toward researchers' attitudes toward bibliometrics (Aksnes and Rip 2009; Buela-Casal and Zych 2012), fewer still have studied actual uses of bibliometrics and their consequences for knowledge production (Rushforth and de Rijcke 2015).

Studies regarding the formalized uses of metrics in research assessments are more common, and a literature looking at practices and effects is gradually emerging (de Rijcke et al. 2016). While acknowledging the importance of these approaches, we suggest that metrics might have even more profound influence when decisions concerning individuals and smaller groups are at stake. For this reason, it is important to engage with the uses of metrics in high-stake contexts, where employing bibliometric indicators might have major consequences for the individual researcher.

Our main focus in this article is the uses of metrics in giving judgments of applicants for academic positions. More specifically, we investigate how bibliometric indicators are used for ranking candidates in two research fields: biomedicine and economics. Based on qualitative analysis of written assessment reports of applicants, issues addressed in

our study concern questions such as: What kinds of bibliometric measures are being used to evaluate candidates for academic positions? In what ways are these measures used? And how are different indicators compared, negotiated, and discussed in forming expert opinions?

Our findings hope to elucidate the different types of bibliometrics used for evaluation purposes, and in doing so open up understanding of how individuals are evaluated. Our selection of disciplines is motivated by an ambition to study fields that both draw on metrics, but which differ in their social and intellectual structure. Drawing on the works of Whitley (Whitley 2000; Whitley and Gläser 2008), we infer that differences in the organization of research fields are likely to have direct consequences for the formation of evaluation practices. The heterogeneity in research practices and publication strategies, as well as agreement on research goals and methods, are some of the factors that are likely to influence the assessment of research among members of the respective communities. The widespread presence of bibliometrics in biomedicine and economics has been widely reported in the scientometric research literature, mostly by way of technical discussions about measures used to evaluate outputs (Graber, Launov and Wälde 2008; Haucap and Muck 2015). The broad coverage of biomedical literature in *Web of Science*, and later *Scopus*, together with the sheer size of the field has contributed to a frequent use of performance indicators in the field (de Bellis 2009; Van Eck et al. 2013). Moreover, recent debates on the supposed crisis in science, where issues regarding how research and researchers are evaluated, have emerged very visibly from the biomedical field (Benedictus et al. 2016). Yet, systematic studies on how research is evaluated in the recruitment of medical researchers are still largely absent.

Economics is one of the larger disciplines in the social sciences, and due to its close connection to the state and the economy, it is also one of the more influential (Maeße, in press). Several studies have also noted that economics distinguishes itself from other social sciences in its hierarchical organization (Whitley 2000; Fourcade et al. 2015), which may also affect evaluation procedures in the field. In addition, publication and citation patterns in economics allow for the use of citation databases on a much more widespread scale than in other social science disciplines (Hicks 2004). Yet, although several studies point to the influence of metrics in these fields, we find little research on how indicators are used to assess and rank individual researchers.

In this study, we relate the complexity of ranking candidates to the heterogeneity of entities being evaluated; all candidates are unique and no single criterion can be used to make judgments. The process of evaluating and ranking candidates can be compared to the valuation of what Karpik (2010) calls 'singularities': unique products that cannot easily be compared—art, literary works, medical doctors, etc.—or valued on a market. Their valuation is therefore dependent on 'judgment devices' to facilitate a uniform ranking of items. Consequently, we propose that bibliometric indicators can be viewed as devices which are used as aids for making decisions when recruiting academic personnel. As such we show how domain expertise in these assessment reports is not so much replaced by bibliometrics, but rather gets redefined through them in quite diverse forms. Based on this analysis, we contribute to an emerging discussion on what is tentatively being called 'citizen bibliometrics' (c.f. Wouters et al. 2015; Leydesdorff, Wouters and Bornman, in press). In recent inquiries the term has been introduced to denote nonprofessional use of bibliometrics by managers and researchers which, importantly, is preferred over more commonly used but pejorative terms like 'amateur-'or 'layman- bibliometrics' (Gläser and Laudel 2007). So far 'citizen bibliometrics' remains an underspecified concept, but to our minds, it carries highly promising normative and descriptive implications, which we will seek to expand upon here. Normatively, it might connote ideas of being a 'citizen' (as in the case of 'citizen science') and part of a collective with certain rights but also duties. Building on the notion of 'academic citizenship' (Macfarlane 2007), citizen bibliometrics implies a certain responsibility, or even care, for one's discipline. As such 'citizenship' relates less to the morality of individuals and more to the ethics of a collective endeavor. While standing for what we consider attractive principles, in our view, the term can also be made to perform analytically useful work as a sensitizing concept, which draws attention toward how reviewers exhibit domain-specific knowledge when using bibliometrics. This is a useful counterpoint to the more technical engagement with the uses of indicators that tend to dominate scientometric discussions concerned with policing 'proper uses' or 'misuses' of indicators. This does not mean we necessarily endorse referees' bibliometric practices as correct usage, but rather it is our intention to use the term to illuminate how reviewers convey what they consider 'good bibliometric practice'. Focusing on how reviewers use bibliometric measures to define and defend notions of excellence in the evaluation of researchers within their disciplines, we suggest that the concept of 'citizen bibliometrics' carries a useful reminder that indicators are always constantly modified, (re)created, and criticized in the contexts of their usage. Thus, rather than consider expertise or responsibility in terms of a priori sets of principles, we intend to focus on how responsibility and academic citizenship play out with respect to bibliometrics in actual evaluation sites. This leads us to argue that while the notion of 'informed peer review' typically rests upon a binary distinction between disciplinary expertise, on the one hand, and knowledge about bibliometric indicators, on the other, a third mode of expertise is important in these evaluations: knowing how and when to deploy indicators in a specific disciplinary evaluation context. Our analysis therefore contributes to discussions about citizen bibliometrics by conceptualizing the knowledgeability of bibliometric judgment devices in a disciplinary domain as an important constitutive element of academic citizenship.

These ideas will be sketched out further over the subsequent pages, which will be structured as follows. To contextualize our findings, the following section examines previous studies on researchers' uses and understanding of bibliometrics. Important concepts which are introduced and developed through our empirical materials are then discussed. Following an overview of the choice of documents used as empirical materials and the methods we used to analyze them, our findings are presented. First, we present an overview of metric uses in the two fields studied, followed by a more detailed analysis of what we call 'the context of bibliometrics', which includes how different types of indicator expertise informed the uses of bibliometrics within these assessment reports. We conclude by discussing how a conceptualization of bibliometric indicators as judgment devices might further our understanding of the concept of 'citizen bibliometrics' and what our proposals might entail for the future study of evaluative bibliometrics.

1.1 Researchers understanding of bibliometric indicators

Scientometricians have criticized the statistical properties of indicators such as the JIF and h-index (Van Leeuwen 2008; Waltman and

Van Eck 2012; Larivière, Lozano and Gingras 2014), and similar criticisms have been directed toward journal ranking and rating lists in economics (Malsch and Tessier 2015; Tourish and Willmott 2015). The importance of these criticisms cannot be underestimated, but our approach more closely resembles *verstehen*-type studies, for instance, those which consider how researchers make sense of indicators. Previous findings indicate that researchers' perceptions of citations and citation-based measures are ambivalent (Hargens and Schuman 1990; Buela-Casal and Zych 2012). For example, respondents in a survey by Aksnes and Rip (2009) stated that they were aware of how citations are used for evaluative purposes, but at the same time, they do not keep track of citations to their own work. Partly, these results could be the consequences of researchers perceiving that to take too much interest in one's own citations is frowned upon. Researchers also form so-called 'folk theories' about citations, where, for example, some claim that being 'trendy' is more important than quality or that the citation rate of the paper is dependent on the status of the author (Aksnes and Rip 2009). In a later survey by Derrick and Gillespie (2013), more than 60% of the respondents reported that they would include bibliometric indicators in an application if they perceived it as advantageous, while an equal percentage of respondents (60%) agreed with the statement that indicators encourage researchers to 'cheat' and 'game' the system.

While acknowledging the importance of researchers' perceptions of bibliometrics, we also believe that these ambivalent attitudes may partly reflect a discrepancy between attitudes toward indicators and their actual use. Hence, this study contributes to an emerging literature studying *uses*—rather than statistical properties or perceptions—of indicators. Taking such an approach, Rushforth and de Rijcke (2015), for instance, showed how biomedical researchers rely on the JIF when making practical decisions on collaborations and publication venue. Our study builds on these more direct approaches toward understanding how researchers use metrics by exploring how indicator uses differ between disciplinary cultures.

1.2 Why analyze reports?

To our knowledge, this is the first study that systematically analyses the use of bibliometric indicators in assessment reports that evaluate and rank candidates for academic positions. As is largely typical of evaluation in higher education and research, expert peer review is the method of choice for ensuring such outcomes. In this particular recruitment, context peer review is performed remotely (Gläser and Laudel 2005) by individuals provided with the same information and asked to assess according to predefined criteria. This is a traditional method of peer review in which peer judgment is an important input in decisions taken by some other agent (Bozeman 1993),[1] in this case appointment committees hired by Swedish universities. Reviewer candidate reports constitute a pertinent empirical resource for exploring uses of bibliometric indicators as judgment devices in research evaluation, as such documents always feature evaluations and (usually) rankings of careers within the Swedish 'academic market'. Moreover, these reports also tend to articulate disciplinary norms for how researchers are evaluated, and deliberations made in these documents have significant influence on individual careers. The study of these documents therefore offers unique insights into the uses of bibliometrics in situations where a lot is at stake, and where indicators are used both for making and justifying complex decisions.

1.3 Recruitment procedures in Swedish academia

Studying assessment reports for academic positions in Swedish academia has two distinct advantages. First, according to 'offentlighetsprincipen' (openness principle), all documentation on decisions made by state institutions in Sweden should by law be accessible to the public. Second, the procedures of external recruitment are fairly similar among institutions for higher education where external assessments, together with interviews and invited lectures by leading candidates, form the basis from which a formal recruitment decision is made. Recruitment procedures vary across national contexts (Musselin 2009), and the system in Sweden (and in Scandinavia at large) is fairly standardized and transparent compared to many other international cases. However, our assumption is that the actual evaluation criteria reflected in these documents are largely dependent on the discipline rather than on national context, and many of the assessment reports are written by referees from abroad. Hence, although our study concerns assessment procedures in a Swedish context, we suggest that our findings raise questions relevant to a broader, cross-national, disciplinary context.

The tradition of using external appraisers, so-called 'sakkunniga', has a long history in Swedish academia stretching back to the late 19th century. This system was introduced to preserve the legitimacy and independence of the university, and gradually, these reports came to play a normative role when defining disciplinary notions of 'scientific quality' (Nilsson 2009). Over time the importance of the external assessment of research quality has lessened somewhat, as other merits such as teaching and administrative skills have been given more weight; yet, skills in teaching are still usually trumped by research merits (Brommesson et al. 2016).

The recruitment procedure in Swedish academia is designed to be impartial and merit-based in that external reviewers assess the candidates; yet, there are many ways in which the recruiting department can influence the process. The broader politics and practices of academic recruitment is indeed a fascinating topic, which has so far only briefly been covered by literature on academic job markets (cf. Musselin 2009). However, in this study we zoom in on one specific part of this process: the assessment of research merits within external expert reports, with a special focus on the use of bibliometrics for assessing candidates in reports during the external peer-review stage of recruiting and promoting.

2. Indicators and rankings as judgment devices

When external referees are assigned the task of ranking candidates, they are called upon to provide order and reduce uncertainty. Their assignment is particularly difficult, as the individuals being evaluated are unique, multifaceted, and complex. Each candidate has unique competencies which cannot be compared directly; the information provided by each applicant is at least partly distinctive, even if general criteria exist, and there are some agreed-upon rules for how the assessment should be performed. Borrowing from economic sociology, we find a parallel between evaluating these candidates and the valuing of unique goods, or what Lucien Karpik (2010) terms *singularities*. A singularity is a good that is unique and not readily compared to other products or services, such as a work of art, a novel, or a medical doctor. The difficulty of assessing singularities makes external support necessary to reach a decision. Customers, or in our case referees, rely upon external support in the form of *judgment devices* that facilitate and legitimate arguments

and decisions. Judgment devices can, according to Karpik, be divided into five main types: *networks*, *appellations*, *cicerones*, *rankings*, and *confluences*. We suggest that two of these, appellations and rankings, are particularly useful for understanding the role of bibliometrics when used in the context of evaluating researchers. Appellations are brands and titles that assign meaning and worth to a product or a group of products. In our case this could involve the brand of a journal (like *Nature*), but could also be a certification (e.g. journals indexed in *Web of Science*) or an origin (e.g. a journal/book published by *Oxford University Press*).

The effectiveness of appellations builds on shared conceptions regarding the identity and quality of a particular label (Karpik 2010: 45–46). In cases where such agreement does not exist an option might be to make use of rankings. Rankings arrange singularities in a hierarchical list based on one or several criteria, and Karpik distinguishes between two different types of rankings: those that build on expert rankings and those that make use of buyers' choices. Expert rankings build on valuations made by domain specialists, and they could take the form of prizes annually awarded or public rankings of universities or hospitals. Buyers' rankings, on the other hand, are determined on the basis of sales of particular products (e.g. 'top 10 lists' of most highly sold products). When consumers draw on judgment devices for decision-making, they agree to trust an external source and do not always understand how the device works or has control over it (Karpik 2010: 46). Eventually, as more trust is invested in these devices, the debate no longer concerns if a single device makes a fair valuation, so much as how different judgment devices stand up against each other. Here evaluation of goods is replaced by the evaluation of judgment devices (Karpik 2010). Although delegation toward judgment devices is not absolute, it is certainly noticeable in our materials, thus prompting us to explore how knowledge and care in using judgment devices manifest in expert judgments about candidates.

We outline the uses of bibliometric indicators as judgment devices in the following ways: as references drawn on to substantiate claims about journal (article) quality in a candidate's CV; listing each candidate's rating on a specific measure without passing explicit comment—thus showing how indicators are used as part of a formal practice of constructing an evaluation report for others to base decisions upon—and combining different judgment devices to cross-validate each other, in supporting a statement about one or more candidates.

3. Material and methods

In very general terms, qualitative research is especially useful for exploring topics where previous literature is either sparse or diffuse. We consider the uses of indicators in evaluation contexts to be one such instance. The analysis of bibliometric uses, which we will present in our findings section, emerged via an interpretive coding framework. To analyze the theme 'bibliometric uses' in the external assessment reports, we followed an inductive content analysis approach, where our results are derived more from the data than an a priori theoretical framework. The approach we took followed the *preparation*, *organizing*, and *reporting* phases set out by Elo and Kyngäs (2008).

In the *preparation phase*, we gathered our empirical materials, namely, external assessment reports from four major universities in Sweden—the University of Gothenburg, Lund University, Uppsala University, and Umeå University—that conduct research both in economics and biomedicine. The assessment reports were produced in recruiting for positions as senior lecturer or assistant professor ($n = 136$) and full professorships ($n = 52$). The definition of economics was quite straightforward, as the field was judged to be equivalent to the Swedish term 'nationalekonomi'. Biomedicine is a much more loosely defined term and therefore we included all specialist positions involving natural science applications in medicine. We collected material from a 10-year period starting in 2005 and ending in 2014 (Table 1). We focused on external reports of applications, which were evaluated in competition, and cases with only one applicant were excluded. Joint statements by several examiners were treated as one report, while independent reports pertaining to one particular case were treated as stand-alone documents. There were common structures to these reports, including a general introduction where the task at hand was to describe each candidate, which then led up to a ranking of applicants. Depending on the number of applicants and the ambition of the reviewer, each of these reports ranges from a couple to over 30 pages.

External assessment reports for academic positions at state-financed universities in Sweden are available to researchers without obtaining permission from the reviewers writing the reports or the candidates being assessed. While both examiners and candidates are probably aware that colleagues or others interested in these processes may read these documents, we decided neither to reveal the identity of either referees nor candidates. Our decision rests on two premises: (1) neither the referees nor the candidates had any opportunity to decide if they wanted these documents to be part of a research project, and (2) we did not find that revealing the identity of candidates or examiners would improve the analysis. Consequently, all reports were coded based on year, field (biomedicine: bio or economics: eco), and university (Lund University: LU, University of Gothenburg: GU, Uppsala University: UU, and Umeå University: UMU). Although it was decided that both authors would be involved in the data organizing process, with only one being a native Swedish speaker (B. H.), the other author (A. R.) was restricted to coding documents written in the English language ($n = 109$). Quotes from Swedish language documents which are displayed in the findings section are based on B. H.'s translations into English.

In the next phase of our inductive content analysis, we sought to *organize* the collected data. This iterative process involved first careful readings and open coding of the texts, followed by rereading and grouping emerging categories into more refined categories. Finally this phase involved 'abstraction', whereby subcategories which fall under the 'main category' (i.e. bibliometric uses) are grouped and refined hierarchically under content-characteristic terms (e.g. JIF, rankings, h-index, etc.). Journal rankings are not a bibliometric indicator in a proper sense, but they are often in part derived from bibliometric indicators and play a similar role as judgment devices in these texts. We therefore included journal rankings and ratings in our study. Although adopting a rather generous and open-ended definition of 'bibliometric use' in the early stages of coding, these subsequent steps allowed us to arrive at a more precise definition of our unit of analysis, restricting 'bibliometric uses' to instances where the JIF, h-index, journal rankings, or citations were explicitly employed with references to numbers in the reports. As such, the final reporting phase of our analysis excludes instances of what we had considered 'bibliometric use' in earlier open-coding rounds, for example omitting categories like 'simple publication counts', or categories derived from general remarks about 'high impact journals' or 'top

Table 1. Overview of studied material

Field	Lund University	Umeå University	University of Gothenburg	Uppsala University	Total
Biomedicine	46 reports	3 reports	22 reports	61 reports	132 reports
Economics	17 reports	4 reports	27 reports	8 reports	56 reports

journals'. Although likely to be inferred indirectly from implicit knowledge of bibliometric measures or rankings, we decided these latter categories were rather too general to be helpful in illustrating distinct 'metric cultures' in biomedicine or economics.

In *reporting* our findings, we present quotes which are illustrative of the most frequently occurring subcategories that emerged in the analysis process. Where it is deemed helpful, we provide simple numerical counts of the incidences of major and subcategories of 'bibliometric uses' in the texts. These are meant to give an impression of how often uses of different metrics are made, not to make statistically generalizable claims about the incidences of bibliometrics in different evaluation cultures.

Based on our sample size and the strict definition of metric use we arrived at, we found there may be some risk in overstating the presence of bibliometric indicators in the context of evaluating candidates for academic positions. For example, a simple count revealed that 82 of 188 reports explicitly used metrics. The value of our qualitative approach is in opening up how metrics are drawn on to substantiate, question, and negotiate specific claims within a familiar yet underexplored research evaluation setting in different disciplinary domains. This enables us to draw (modest) conclusions and formulate further research questions about the role of bibliometric indicators as 'judgment devices' in research evaluation and to flesh out our own tentative insights in relation to the emerging concept of 'citizen bibliometrics'.

4. Findings

4.1 Using metrics in biomedicine and economics

In this section, we compare the uses of different metrics across the fields of biomedicine and economics. From reports in which indicator usage was found, we observed some similarities across disciplines, for example in indicators being introduced without hesitation, and with an assumption that they directly reflect the scientific ability of the applicant:

> Scientific skill has been judged based on scientific publications and citations to these publications registered in Scopus (www.Scopus.com) as well as h–index which highlight the quantitative influence of the author or scientific impact. (Bio GU 2013–5, p. 1)[2]

> A bibliometric analysis was carried out to assess the scientific production and even more importantly, the real scientific impact of each applicant. (Bio LU 2014–3, p. 5)

In the latter quote, it is not even that metrics 'highlight' certain aspects, but it is said to represent *real* scientific impact, implying more qualitative descriptions may result in inaccurate assessments of applicants. The use of metrics is also motivated by indicators being 'unbiased' (Bio UU 2008–2, p.1; Bio UU 2012–11, p.1) and therefore providing fairer assessments of candidates. Although such strong sentiments advocating bibliometrics are rare, we found that metrics were often presented in a neutral or positive tone across both disciplines. Another explanation used to validate uses of metrics was the sheer volume of information that reviewers have to take into account:

> Expert appraisals can be rather long tomes, and so I have attempted to utilize tables to compact the information provided by the candidates and available from other sources, such as the Web of Science. (Bio UU 2012–4, p. 1)

Resonating with Karpik's account of judgment devices, this quote suggests reviewers face problems with *excess*, both in the sense that there are several possible candidates and an abundance of information regarding these candidates. For the individual, this may result in overload, a situation which can be solved either through the reduction of excess or through redefining excess (Abbott 2014). We have observed that the uses of bibliometrics equates to a reactive strategy for reducing excess by '… hierarchizing and concentrating one's attention at the top end of the hierarchy' (Abbott 2014: 18). In many cases reviewers point to an abundance of quality candidates, rather than scarcity. In this situation ranking and rating devices offer a means of sorting between 'good' and 'very good' candidates.

These arguments for using bibliometrics possibly point toward epistemic practices of evaluation in biomedicine in which expert judgments are legitimated by way of mechanical, standardized, 'objective' indicators (c.f. Porter 1996). This resonates also with Lamont's (2009) observation that realist commitments toward objectivity in academic peer review often stem from epistemological traditions in which the evaluators are embedded as researchers (while peers from social constructionist fields in contrast feel more comfortable with the 'subjective' dimensions of peer-review processes and are skeptical of bibliometrics). Although in a broad sense these observations appear tenable, as we will show, bibliometric use in several respects differs considerably between biomedicine and economics.

A little less than half of all assessments in biomedicine made explicit use of bibliometric indicators (58 of 132 reports, 44%), and looking at reports regarding professorships revealed that there was little difference in explicit metric use depending on the position advertised (26 of 52 assessment reports concerning professorships used indicators). Among the more frequent indicators used in biomedicine were the h-index (26 reports, 20%) and the JIF (23 reports, 17%), while straight or adjusted citation counts were found in 38 biomedical reports (29%).

The proportion of reports in economics that use bibliometric indicators or journal rankings in assessing applicants is almost exactly the same as in biomedicine 24 of 56 (43%). Straight or adjusted citation counts were the most common indicator in economics (13 reports) with journal rankings also being frequently used (11 reports). JIF was given in nine reports, while h-index was used in only three cases. These figures are not intended to support claims for statistical generalizability, but provide nonetheless a useful précis of how frequently different indicators featured across our materials.

Comparing the two fields, we found that reviewers in biomedicine tended to use JIF scores and the h-index, while journal

rankings—which were not used at all in biomedicine—form a tradition in economics. The use of JIF in biomedicine could be interpreted as a form of rating, or in Karpik's vocabulary *appellations*. Here the JIF becomes a brand in the sense that researchers refer to journals with a high impact factor as specific types of journal: the JIF becomes a 'stamp of quality'. This type of judgment device is most effective when evaluation criteria are well recognized and agreed upon, as is often the case with research fields where there is considerable agreement regarding standards and technical procedures. Moreover, appellations appear as especially useful in fields like biomedicine in which no specific group dominates when defining criteria for scientific quality and where several groups influence the field in terms of funding and employment (Whitley 2000). Consequently, the form of rating, or 'branding', used in biomedicine is effective due to a general agreement on research procedures and evaluating criteria. Economics, on the other hand, distinguishes itself from other social sciences in having a low degree of strategic uncertainty, and it is also characterized by a rather high degree of mutual dependency, a type of organization that Whitley (2000) describes as a 'partitioned bureaucracy'. Economics is distinctive for the hierarchical structure of the discipline, which is upheld by a reputational élite through its influence over the training of new economists, the communications system, and access to resources (Coates 1993: 42; Maeße in press). The voluminous production of rankings—of institutions, journals, and scholars—in economics appears illustrative of what Fourcade, Ollion, and Algan (2015) claim is a predilection for hierarchies within the field. The material under study here, in which no less than five different rankings are used, appears to resonate with these more general characteristics. The extensive use of journal rankings also suggests that the most relevant literature in economics is found in a distinctive and rather small set of key journals. Thus, the comparatively insular nature of literature in economics, which is confirmed by bibliometric studies of citation patterns (Fourcade, Ollion and Algan, 2015), is a further factor which might help to contextualize the popularity of journal rankings within the field. In sum, our findings appear consistent with Whitley's account of both economics and biomedicine as fields featuring characteristics which would seem conducive for the extensive use of bibliometric indicators in recruitment procedures. However, already at the outset, we also identify differences across these fields in the type of bibliometric indicators used, and discipline-specific characteristics. These differences will now be made further visible in a detailed description of indicators as judgment devices.

4. 2 Uses of indicators as judgment devices

4.2.1 The Journal Impact Factor

In this section we show how indicators have come to assume different meanings and acquire different uses as judgment devices within our materials. The JIF, built on an idea forwarded by Gross and Gross in 1927 (Gross and Gross 1927), later realized by Eugene Garfield and incorporated as a feature in the Science Citation Index (Garfield and Sher 1963), is one of the most popular and at the same time most criticized bibliometric indicators (Archambault and Larivière 2009). By calculating the average number of citations per article in a journal, the JIF is said to indicate the 'impact' and relative standing of a periodical. In our materials we see how JIFs are used to establish orders and values among publications in the reports. A common practice is to attach JIFs in the text to support statements on journal quality. This is often done in a 'neutral' reporting type fashion with the JIF being given in a parenthesis after the name of the journal—almost like a reference in scholarly text to substantiate a statement, only in this case relating to the importance of a journal:

> . . . but it is a bit bothersome that many of the recent publications that XXXX has been principal investigator on are found in more narrow journals, for example Scandinavian J Immunol. (Impact approx. 2.3). (Bio GU 2012–2, p. 2)[3]

> His CV includes 20 peer reviewed publications in journals such as Physica D, Studies in Nonlinear Dynamics and Econometrics (Impact factor 0.593), European Journal of Health Economics, Applied Economics (Impact factor 0.473), (. . .) European Financial Management (two articles, Impact factor 0.717), Journal of Economics and Business, and Energy Economics (Impact factor 1.557). (Eco UU 2009–2, p. 2)[4]

Similar uses are found in several reports where JIF is given as supportive evidence of journals having a good reputation, or, as the illustration below makes clear (Table 2.), it can be presented in a table showcasing journals, their impact factor, and the number of publications published in the same 'top journal' for each candidate.

Another common use is to indicate a scoring interval (e.g. ranging from 4 to 7) of the JIF of journals where the papers of an applicant have been published (Bio GU 2006–1; Bio UU 2013–7). Such scales are taken to reflect not only the ability but also the ambition of the researchers in question. Aiming, and subsequently succeeding, in publishing in high impact journals signals a resourceful and successful applicant capable of overcoming peer review in journals with high rejection rates. The JIF is used as an obvious shorthand for a journal's reputation and by association the standing of a candidate. However, in cases where the interval is broad—for example stretching from 0.5 to 26, or from ordinary to highest quality (Bio UU 2008–4)—such numbers carry little meaning and have to be supplemented by other judgment devices. A few reviewers take this a step further by aggregating JIFs and then coming up with averages or medians of impact factors (Bio LU 2005–6; Bio UU 2012–4).

JIF scores can also be used for setting a standard or a benchmark. This is illustrated in a report claiming that most journals (in which the applicants have published) have 'a JIF over 4' (Bio UU 2012–7) or by counting the number of papers published in journals with an impact factor of 5 or better (Bio LU 2005–7). The magic number for a publication to be regarded as of high quality in biomedicine seems in many cases to be around 3 (Bio UU 2011-3) or slightly higher:

> Many of original papers appeared in excellent quality journals and nearly two third [sic] of them were in journals with impact factor greater than 3. (Bio LU 2014–4)

Table 2. Table presenting journals, impact factors of these journals, and the number articles published by a candidate (Bio LU 2006–3, p. 4)

Journal	Impact factor	Number
Circulation	10.94	3
Eur Heart J	7.92	15
JACC	9.7	4
Ann NY Acad Sci	1.93	1
Br Heart J	3.7	5
Pacing Clin Electrophysiol	1.56	7

In economics, where citation frequencies are generally lower, we find statements claiming that JIFs of around 0.7–0.9 are normal for average field journals (Eco UU 2009–3), while others suggest that journals with impact factors over 0.5 are highly ranked (Eco GU 2010–1). Hence, what is to be considered as a high impact factor in the field of economics is not evident, and generally, we find that impact factors in economics are accompanied by qualifying statements, which help the reader to evaluate the score which the judgment device produces.

In biomedicine especially, a JIF of a certain magnitude functions as a benchmark for what is to be considered a high-quality journal, and eventually this 'stamp of quality' also serves as a device used for making judgments on the merits of individual researchers. The statement that specific journals have 'high impact'—which explicitly or implicitly is derived from the long tradition of using the JIF in biomedicine—moves away from the context in which these numbers are produced, and becomes a 'fact' of its own. The JIF functions as an appellation by assigning value to the journal. Eventually the impact factor becomes part of the 'brand' of the journal. Following this line of thought, we would suggest that the influence of impact factors goes much further than their actual use, as they come to form a whole way of thinking and vocabulary for discussing quality.

4.2.2 The h-index

Whereas JIFs are often given in sections of the report where specific research contributions (in the form of journal articles) are discussed, the h-index is often included in general descriptions of the applicants. Invented by Jorge Hirsch—a physicist and not a 'professional bibliometrican'—the h-index is a very well-known attempt to come up with an indicator that reflects both the quality and the quantity of publications produced by an author. In short, if a scholar has an h-index of h, it means that she has published h publications which have been cited at least h times (Hirsch 2005). We found that the h-index could form the subject of discussion in the texts, but to a lesser extent than the JIF does. Unlike the JIF, the h-index was often given as a stand-alone 'fact' about the applicant:

> XXXX publishes in good to very good journals including Plos Genetics, FASEB J, and Mol Biol Cell. H–factor=18. (Bio GU 2013–9, p. 3)

In other reports the h-index was included, together with other 'details' such as date of birth, current position, etc., to provide background information on the candidates (Eco UMU 2009–1, p. 8; Bio LU 2008–5). We also found examples where the h-index was introduced in the narrative using 'screen dumps' from *Web of Science* (Bio LU 2013–2). Here the h-index becomes closely connected to the person being evaluated; it helps to identify and characterize not only the scientific production but also the applicant as such. Giving the h-index alongside other basic information also appears to heighten the importance of the measure as a necessary background fact, which is given before the actual narrative begins. Thus introducing metrics into the text as a 'mere formality' exhibits one way in which particular judgment devices can become taken for granted. The h-index score is not only commented upon but is presumably expected to 'speak for itself' as an indicator of the candidate's standing in relation to individuals with whom she is being compared. There is thus an expectation that the persons reading the report will want to know this score and in part base their decisions upon it.

The h-index could be seen as an attempt to summarize a whole career in one single measure, and in some reports, the h-index is represented as an almost magical number that can be used to characterize and grade a researcher—denoted by its stand-alone positioning within a report with no other textual information surrounding it to give an explanation.

The totalizing effects of using the h-index is illustrated by these numbers being hard to ignore once they are given, and there are several reports where they tend to play a decisive role in the referee's recommendations. The most evident example is found in Bio UU 2014–1 where the h-index of each candidate corresponds almost perfectly with the assessment and rankings made. Upon closer inspection, the final judgment made of 23 candidates for a professorship at Uppsala University (Bio UU 2014–1) aligns with the h-indexes of the candidates (Table 3).

Although certain judgment devices have become very important for evaluation in many fields, no one device completely dominates. In our materials, for example, the h-index is quite often combined with other indicators, such as straight citation counts or JIFs (Bio LU 2013–2; Bio UU 2012–4).

This section suggests that forced decision-making situations such as these lend themselves to the use of judgment devices, as reviewers must recommend someone or something from a range of 'commodities', all with unique multidimensional qualities ('singularities') (c.f. Karpik 2010). The authority and expertise of the reviewers are exercised in different ways across our materials. Whereas in some cases expertise is implied via the qualitative judgments of reviewers and are expected to carry weight in their own right, we have also made visible in this section and throughout the article that expertise can be displayed through using various bibliometric judgment devices. In the next section we consider further the dimensions of expertise demonstrated by uses of metric indicators in this evaluation context, and in doing so revisit previous characterizations of uses of bibliometrics as 'amateur', proposing instead the term 'citizen bibliometrics'.

4.3 Citizen bibliometrics

In this article we deliberately use the term 'citizen bibliometrics', first proposed by Wouters et al. (2015), as opposed to 'amateur bibliometrics' or 'nonprofessional use'. Rather than thinking of indicators as marking an 'outsourcing' of judgment and expertise to more mechanical 'objective' procedures, the term citizen bibliometrics allows us to concentrate on how indicators appear to be redefining what is meant by expert judgment in particular research fields and institutional settings. What we see in our materials is indeed that some reviewers are quite knowledgeable about bibliometric indicators and their shortcomings. For example, part of demonstrating expertise in these contexts is to cite the limitations of bibliometric indicators, as well as demonstrating awareness of which indicators to deploy or not deploy in evaluating research outputs from their

Table 3. h-index and recommendation for professorship (Bio UU 2014–1)[5]

h-index	Recommendation by the referee
0–14	Not qualified/not eligible
15–20	Borderline qualified
21–25	Qualified
26–33	Fully qualified

own field. These qualifications and selective uses demonstrate not only some technical knowledge but also an informal responsibility to and care towards one's disciplinary norms and standards.

Despite their wide acceptance, external assessors might still find it necessary to explicitly discuss the use of bibliometric measures. There are, for example, cases where the general assumptions about JIFs are problematized when making judgments on applicants from different fields:

> Nuclear medicine journals do not have really high impact factors (not like e.g. Lancet and Nature having impact factors >20). The best journals focused on nuclear medicine most often has impact factors <10). (Bio LU 2014–6)

Interestingly, this justification is found in a report that does not make explicit use of JIFs, but the expert in this case, well aware of the readers' tacit understanding of the report, wishes to make this difference explicit. Besides knowing the limitations of certain indicators, some examiners make judgments on which ones to use:

> I do not use citations as they are unreliable when citation windows are short and unevenly distributed. When it comes to journal impact the results are dependent on the measures used to a considerable degree. (Eco GU 2013–1, p. 8)[6]

As well as questioning the reliability of indicators, some also discuss the validity of indicators:

> Impact measures of this kind are inexact and should not, in our view, be relied on for a detailed ranking of research achievements (it could be described as 'a scale that can distinguish an elephant from a rabbit but not a horse from a cow'). However, the ability to publish in influential and selective journals is important in the Economics field and therefore such rankings provide useful information on the quality of research. (Eco GU 2012–2, p. 1)

In this case the reviewer hesitates when introducing metrics and questions if metrics can be used to rank candidates, especially in cases where they have similar merits. Yet, the reviewer still finds them useful, as they disclose an ability to publish in highly ranked journals—a skill that is valued in economics. The gist of the argument seems to be that the use of these indicators is justified because it is an accepted method for assessing research quality within the field, and not because the measures as such are very sophisticated. This sentiment is also evident in the following quote from an assessment report in economics:

> The ability to publish in influential and selective journals is important in the field of economics, which means that the ranking of journals influences the assessment of research quality. (Eco UMU 2012–2)[7]

In our material we find that the question of whether metrics should be used at all is supplemented by a discussion on how indicators are best employed for judging candidates. There are indeed some examiners that are hesitant about using bibliometric indicators, as in a case where the distinction between 'assessing the candidates' and 'consulting' a database is emphasized:

> After a first review of the applications I consulted 'ISI Web of Knowledge' for citations—and publication numbers. This has not changed my evaluation or my assessment of the candidates. I am aware that professor xxx has included these analyses (in extenso) in his evaluation, and therefore they are not included here. (Bio LU 2011–1, p. 1)[8]

Notably, this examiner not only desists from the temptation of using metrics when performing his evaluation but also subtly criticizes this colleague for relying too much—'in extenso'—on metrics.

Besides discussions on the applicability of indicators more generally, we also see that more technical aspects are discussed such as the length of the citation window (Eco GU 2008–2, p. 1, Bio UU 2012–4), or the usefulness of specific databases:

> Apparently, it takes time to make an impact in World of Knowledge [sic. Web of Science] and this limited information source is not useful for discriminating between applicants. An alternative, with a larger coverage, is Google Scholar and here we find rather large differences. (Eco LU 2010–3, p. 12)[9]

Comparisons are not only made between different bibliometric data sources but also concerning specific indicators used. External examiners negotiated the results of bibliometric measurement by introducing other metrics that question established indicators. In other words, they question one judgment device by introducing another. For example, one examiner touches upon the relation between the impact of a journal and that of an article published within the same journal:

> It might very well be that a highly cited article in a low ranked journal should be given a higher value than a rarely cited article in a highly ranked journal. (Eco GU 2008–4, p. 2)[10]

This argument does not dispense altogether with the notion that journal rankings have value, but importantly suggests that the meaning of such indicators can and should be taken simply as one possible tool for informing merit. There are also examples of different indicators being juxtaposed to give a more complete overview of a candidate, as in this example where several numbers (published papers, authorship, max citations, and h-index) are given:

> Of 44 published papers she is 1st author on 12 and senior author on 20. She has a surprisingly low citation rate, albeit with a high h–index (Max citation <60 in 2010, h–index = 17, Web of Science). (Bio UU 2012–11, p. 8)

Reviewers might also reflect on problems of data coverage or other crucial issues such as author ambiguity, i.e. the problem of tying publications to a specific author (Bio UU 2012–9).

Introducing time as a factor for contextualizing publication numbers and citation scores is another way of negotiating results. Time in terms of years as an active researcher is often introduced as an issue to consider. In one assessment report in economics, this is done by introducing all applicants with their names tightly followed by 'Years out' and 'citations': 'Years out: 14, Citations 3328' (Eco GU 2014–2, p. 1.) This information is given before any descriptive text, and the numbers are highlighted in bold by the reviewer to reflect their significance. Overall, it is common that indicators are contextualized or even adjusted in relation to time, especially in cases where there are major differences in the professional age of applicants. Temporal aspects may also be considered when future impact is considered, and expectations may influence judgments:

> He has the lowest citation record among the three applicants (151 citations and h–factor of 7, according to Harzing's PP ranking) but looking at the REPEC list of citations he has many recent citations, and my assessment is that his publication and citation profile will be very positive. (Eco LU 2011–8, p. 1)

Table 4. Comparison of candidates (columns) according to different indicators (rows) by one reviewer (anonymized) (Bio UU 2012–4, p. 5)

	Candidate A	Candidate B	Candidate C	Candidate D	Candidate E	Candidate F
Number of published papers/reviews	44/5	26/1	5/0	34/1	9/3	47/3
First/last author[1]	6/11	7/8	3/0	10/1	4/0	18/10
Other senior author[2]				Author X (27/34 papers)	Author Y (6/9 papers)	Author Z (37/47 papers)
Selected papers[3]:	10	10	4[6]	17 (+1 poster)	12	19
Median (range) i.p.	2.48 (2.11–4.13)	2.74 (1.33–31.20)	3.34 (1.86–8.28)	4.19[7] (1.52–35.53)	9.68[8] (5.53–11.66)	3.18 (2.52–9.68)
Median (range) citation rate[4]	9.5 (3–83)	12 (4–81)	22 (6 and 38)	18.5 (5–58)	26 (650)[8]	22 (3–226)
Estimated h-index[5]	16	10	2	14	10	20

This quote exhibits knowledge about the reliability of database sources within the field of economics Research Papers in Economics (RePEc), combined with working assumptions about what constitutes an acceptable number of citations within a certain time window, which a professional bibliometrician working outside of the field of economics would not be able to judge.

Finally, we also see several examples of examiners that bring together a range of indicators with the purpose of comparing them with each other. This is usually presented separately from the text in the form of tables. One of the most ambitious examples of this practice is found in Table 4.

This table introduces the section 'scientific merits' and is accompanied by eight footnotes, which describe and compare these numbers. There are several clues pointing to whether an examiner is accomplished in using bibliometrics for evaluations. Indicators are introduced and explained—and in the case of the h-index even referenced—and detailed considerations are made. For example, in footnote 7, it is stated that: 'The number is lower (2.44) for the six papers where she is first author, but the citation rate, in many ways a more important measure, is not much different' (Bio UU 2012–4, p. 5). Here two key qualifications are made: that citations are more important than impact factors and that authorship order is important to consider. Moreover, it is interesting to note that the median (and range) of citations and impact factors are given rather than the average, a practice that seems advisable from a statistical perspective, as citation scores are generally highly skewed. In this example, several judgment devices are brought in, and their strengths and weakness are scrutinized in an informed discussion.

Another means of demonstrating expert knowledge of indicators is to refer to well-known critiques of mainstream indicators. Reviewers in several instances were aware of the major criticisms of the h-index, including its dependence on the academic age of the candidate (making it an indicator of age as much as one of impact). Several of the examiners were aware of this weakness, and some even came up with ways of solving this issue:

> It could be worthwhile to compare the bibliometric scores for the three strongest candidates. Their h–index are 9 (xxxx), 13 (yyyy) and 14 (zzzz). Their academic careers are of different lengths which makes it interesting to study h–index divided in years after PhD–defence (minus parental leave): yyyy 0.68; zzzz 1.56 and xxxx 0.9. However, zzzz is not senior author of these publications which yyyy and xxxx are on some of the publications contributing to their h–index. (Bio UU 2010–2)[11]

The detailed and knowledgeable use of metrics shown in these examples points to the fact that many, although by no means a majority of, examiners have considerable skills in presenting, explaining, and contextualizing bibliometric data. We would therefore suggest that examiners in some of our documents emerge as experts in three roles: (1) as domain experts (2) experts on strengths and limitations of metrics, and (3) experts on how metrics are used and valued within their field. Furthermore, our findings suggest that a move from evaluating publications (or products in Karpik's vocabulary) to evaluating judgment devices (bibliometric indicators) is evident in some reports. However, in most cases we find that judgments on actual content and the significance of scholarly publications are combined with judgments on the indicators used to evaluate these publications. Expertise is demonstrated through deploying a given judgment device that is deemed appropriate for evaluations in respective epistemic communities, and through mediating between alternative judgement devices. As such, we envisage that the rather empty label of 'citizen bibliometrics' may help to evoke some of these expert dimensions which are seldom brought forward in discussions about 'amateur bibliometrics'.

5. Discussion

In this article, we have proposed that a fruitful way of understanding the attractiveness of bibliometric indicators is in terms of their ability to help form decisions in situations where the quality of a work, or an individual's worth, is difficult to assess. Accordingly, indicators can be understood as *judgment devices* in helping referees to reach a decision in a peer-review context where there is a glut of suitable candidates. In competitive professional environments characterized by 'credential inflation' (Collins 1979), judgment devices play important roles in making distinctions between singularities (Karpik 2010). What, however, does the use of this concept and our empirical findings suggest about the character of 'citizen bibliometrics'? In this section we will outline the contribution of this article, and expand on why the concept of citizen bibliometrics might be more appropriate as both an analytic and normative category than 'amateur bibliometrics'.

Our findings show that the use and development of bibliometric indicators are not restricted to professional bibliometricians, but these measures are, to a considerable degree, also discussed, developed, and refined by other groups. While acknowledging the great importance of knowledge about how indicators are constructed, we want to emphasize that 'citizen bibliometrics' should not be considered as simple use or misuse of indicators developed by bibliometric experts. On the contrary, we find the efforts of developing more advanced indicators—over 100 different indicators of impact have so far been developed by the bibliometric community (Wildgaard, Schneider and Larsen 2014)—are largely ignored when used in the context of assessing individuals. Simple and well-established indicators, like the JIF and the

h-index, are preferred. The reason is probably not only that these indicators are readily available and relatively easy to calculate but also because they are well-established tools of evaluation within some disciplinary communities. Moreover, our study shows that domain specialists quite often possess considerable knowledge about these measures, and they are aware of limitations and field differences when using indicators. In contrast to professional bibliometric experts, they also know which indicators are valued and recognized within their own field. Clearly, these examiners are not amateurs in using bibliometrics—in fact they are actually paid to conduct these assessments—meaning clear-cut distinctions between expert or amateur and professional or nonprofessional use are difficult to make based on our material. Moreover, the suggestion put forward by Gläser and Laudel (2007: 118) that reviewers 'simply trust the numbers', and that this undermines independent peer review, is not supported by our findings.

Often qualification statements were provided, demonstrating reflexivity about the deployment of indicators. Such qualifying arguments also imply the belief that others do indeed use these systems (wrongfully) as decontextualized representations of excellence or quality. Reviewers here justify the uses of indicators and their own expertise in using them *by citing their limitations*, a rhetorical strategy which is often employed also by experts of advanced bibliometrics. Again, this suggests common distinctions between expert and amateur bibliometrics may be too sharply drawn, at least insofar as they do not account for the intermediary modes of expertise we have reported in this article. It is, we believe, these 'shades of grey' that populate evaluative bibliometric practices, and which therefore require further attention if the uses of bibliometric indicators in research evaluation are to be better understood and theorized. To do so, we suggest that directly observing indicator uses is crucial.

Our analysis has led us to suggest that comparisons of indicators as judgment devices across disciplines are productive and even crucial for understanding the influence of bibliometric measures as technologies of evaluation. The giving of h-index or JIF as proxies for impact will mean something different in biomedicine compared to economics, and these numbers are treated quite differently depending on the discipline in question. While bibliometric indicators are used across both disciplines, we find that indicators serving as ratings (e.g. the JIF), or what Karpik (2010) calls appellations is an important part of evaluation in biomedicine, while journal rankings are more popular in the field of economics. In turn, we have argued that differences in the social and intellectual structure of these fields are significant when analyzing evaluation procedures. For example, the organization of economics facilitates a further use of journal rankings, which then accentuates and supports arguments about a preoccupation with reputational hierarchies. The importance of the JIF in biomedicine, which in many ways is also firmly integrated in the production of knowledge (Rushforth and de Rijcke, 2015), relies on rather standardized evaluation procedures. These are, in turn, dependent on widespread agreement on methods and priorities in the field. Overall, the organization of biomedicine and economics, as well as the publication practices of these fields, allows for rather extensive, although not always warranted, use of bibliometric indicators. The idea that disciplinary differences are important for understanding the consequences of bibliometric evaluation is well established, but we suggest that combining a contextualized understanding of bibliometric indicators as judgment devices along with a framework for characterizing the organization of specific fields, opens up greater potential for more detailed analyses of bibliometric use. Such an approach also destabilizes current distinctions between expert and amateur bibliometrics at the same time, as it questions the traditional juxtaposition of (pure) peer review and mechanized indicator use. In fact, these types of evaluation intermingle in our material, and domain expertise and bibliometric expertise cannot always be neatly separated.

In our case, we deem that disciplinarity is one important context of bibliometric use, but national differences and institutional settings might be other important factors to consider for future research on the use of metrics. Furthermore, while our focus has been on the use of metrics, we suggest that evaluation reports for academic positions—a rich but understudied material—may bring insights into how careers are assessed and how authorship is valued. Moreover, an analysis of the applications that are assessed (the documents provided by candidates for academic positions)—and not only the evaluation reports as in this case—might give further insight on the use of metrics in these settings. Efforts, like Nilsson (2009), to study these documents over time might also lead to insights on how specific evaluation practices have evolved. For example, the establishment of online platforms that claim to measure 'impact' might result in indicators, such as the '*ResearchGate* score', becoming further integrated into the evaluative practices of academic fields. In our material *ResearchGate* is mentioned once (Bio UU 2013–2), but due to the popularity of *ResearchGate*, and similar platforms, use of such measures is likely to increase (Hammarfelt, de Rijcke and Rushforth 2016). Finally, while biomedicine and economics are both large and influential fields where extensive use of bibliometric indicators has already been reported in the literature, we also see a need for extending our approach to fields where the use, and possibly knowledge about bibliometrics, is less widespread. For example, 'citizen bibliometrics' in the humanities, apart from being less common, would probably be manifested in rather different ways.

What use then might comparative studies of indicators as judgment devices across different fields be in understanding ongoing debates on evaluative bibliometrics? One aspect highlighted in this article is that it can help to expand assumptions about what constitutes expertise when bibliometrics start to become embedded into evaluation contexts. One logical development in any context where judgment devices are used for reaching decisions is that evaluators start to assess and compare the devices used to make judgments and not the entity being evaluated in the first place. In short, they become experts in judgment devices as much as experts on the specific topic at hand. This is one component of citizen bibliometrics we would like to put forward. The expertise of such reviewers is not necessarily simply in their research specialty (as implied by notions of 'informed peer review'), but in mediating between their own epistemic cultures of evaluation and knowledge production, on the one hand, and the affordances and limitations of specific bibliometric indicators, on the other. Thus, it is knowing *how* and *when* to deploy indicators which should be considered the marker of expertise in such evaluative contexts. We suggest the term 'citizen bibliometrics' is a more inclusive and generous means of conceptualizing evaluative expertise than 'amateur bibliometrics'. Moreover, the term 'citizen bibliometrics' also connotes an ethical and caring dimension, i.e. the responsibility to apply indicators in manners which do not damage the community served by the evaluator.

Funding

This research was supported by the Swedish Research Council (grant number 2013-7368) and Riksbankens Jubileumsfond: the Swedish Foundation for the Social Sciences and Humanities) (grant number SGO14-1153:1).

Acknowledgements

The authors wish to thank Fredrik Åström, Elena Maceviciute, Gustaf Nelhans, Ola Pilerot, and Ludo Waltman, as well as the reviewers, for valuable and useful comments, and also to Frances Hultgren for copy-editing of the final manuscript.

Notes

1. Other examples of traditional peer review include journal peer review in which the executive decision on accepting or rejecting a manuscript is taken by an editor based on information produced by remote referees.
2. Original (Swedish): 'Den vetenskapliga skickligheten har bedömts baserat på vetenskapliga publikationer och publikationernas citeringsgrad dokumenterad i Scopus (www.Scopus.com) samt h–index som belyser författarens kvantitativa påverkansgrad eller vetenskaplig genomslagskraft'.
3. Original (Swedish): 'Många publikationer är i goda tidskrifter, men lite bekymmersamt är att många av de som xxxx är huvudansvarig för på senare tid återfinns i lite smalare journaler som t.ex. Scandinavian J Immunol. (Impact ca 2.3)'
4. Original (Swedish): 'Hans CV inkluderar 20 publikationer i tidskrifter med referee förfarande såsom Physica D, Studies in Nonlinear Dynamics and Econometrics (Impact factor 0.593), European Journal of Health Economics, Applied Economics (2 st, Impact factor 0.473) (…) European Financial Management (Impact factor 0.717), Journal of Economics and Business, och Energy Economics (Impact factor 1.557)'.
5. There is only 1 case of 23 assessments where an assessment diverged from this pattern—a researcher with an h-index of 16 which was deemed as 'not qualified'.
6. Original (Swedish): 'Jag använder inte citeringar eftersom de inte är pålitliga när citeringsfönstren är korta och ojämnt fördelade. När det gäller tidskriftsimpact beror resultaten i ganska stor utsträckning på vilket mått som används'.
7. Original (Swedish): 'Förmågan att publicera i inflytelserika och selektiva tidskrifter är viktig i nationalekonomifältet, vilket innebar att rankningen av tidskrifter påverkar bedömningen av forskningens kvalitet'.
8. Original (Danish): 'Efter den første gennemgang af ansøgningerne har jeg konsulteret "lSI Web of Knowledge" citations– og publiceringstal. Det har ikke ændret min vurdering eller rangordning af ansøgerne, men blot understreget betimeligheden af den primære vurdering. Då jeg er vidende om, at professor xxx inkluderer de nævnte analyser (in extenso) i sin vurdering, er de ikke medtaget her'.
9. Original (Swedish): 'Uppenbarligen tar det tid att göra avtryck i World of Knowledge och för att diskriminera mellan de sökande blir den lilla informationskällan inte så användbar. Ett alternativ som har bredare träffyta är därför Google Scholar och här framgår relativt stora skillnader'.
10. Original (Swedish): 'Det kan har mycket väl vara fallet att en ofta citerad artikel i en lågt rankad tidskrift ska varderas högre än en sällan citerad artikel i en högt rankad tidskrift'.
11. Original (Swedish): 'Det kan vara värdefullt att först jämföra en bibliometrisk parameter för de 3 starkaste kandidaterna. De tre starkaste kandidaternas h–index är 9 (xxx), 13 (xxx) och 14 (xxx). Forskarkarriärernas längd efter disputation varierar mellan de tre kandidaterna, varför det är intressant att studera h–index dividerat med antal år (minus föräldraledighet) efter disputation: xxx 0.68; xxx 1.56; xxx 0.9. xxx är dock inte senior författare på de aktuella uppsatserna, vilket xxx och xxx är på en del av de uppsatser som bidrar till deras h–index'.

References

Abbott, A. (2014) 'The Problem of Excess', *Sociological Theory*, 32/1: 1–26.

Aksnes, D. W., and Rip, A. (2009) 'Researchers' Perceptions of Citations', *Research Policy*, 38/6: 895–905.

Archambault, É., and Larivière, V. (2009) 'History of the Journal Impact Factor: Contingencies and Consequences', *Scientometrics*, 79/3: 635–49.

Benedictus, R., Miedema, F., and Ferguson, M. W. (2016) 'Fewer Numbers, Better Science', *Nature*, 538/7626: 453.

Bozeman, B. (1993). 'Peer Review and Evaluation of R&D Impacts', In: *Evaluating R&D impacts: Methods and practice*, pp. 79–98. Dordrecht: Springer.

Brommesson, D. et al. (2016). *Att möta den högre utbildningens utmaningar*. IFAU-rapport, 4

Buela-Casal, G., and Zych, I. (2012) 'What do the Scientists Think about the Impact Factor?', *Scientometrics*, 92: 281–92.

Butler, L. (2007) 'Assessing University Research: A Plea for a Balanced Approach', *Science and Public Policy*, 34/8: 565–74.

Coates, B. (1993) *The Sociology and Professionalization of Economics: British and American Economic Essays*, Vol II. London & New York: Routledge.

Collins, R. (1979). *The Credential Society*. New York: Academic Press.

de Bellis, N. (2009). *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham: Scarecrow Press.

de Rijcke, S. et al. (2016) 'Evaluation Practices and Effects of Indicator Use–A Literature Review', *Research Evaluation*, 25/2: 161–9.

Derrick, G. E., and Gillespie, J. (2013). '"A Number You Just Can't Get Away From": Characteristics of Adoption and the Social Construction of Metrics use by Researchers", In: Hinze, S. and Lottman, A. (eds) Berlin: *Proceedings of the 18th International Conference on Science and Technology Indicators*, pp. 104–16.

Elo, S., and Kyngäs, H. (2008) 'The Qualitative Content Analysis Process', *Journal of Advanced Nursing*, 62: 107–15.

Fourcade, M., Ollion, E., and Algan, Y. (2015) 'The Superiority of Economics', *The Journal of Economic Perspectives*, 29/1: 89–113.

Garfield, E., and Sher, I. H. (1963) 'New Factors in the Evaluation of Scientific Literature through Citation Indexing', *American Documentation*, 14/3: 195–201.

Gläser, J., and Laudel, G. (2005) 'Advantages and Dangers of 'Remote' Peer Evaluation', *Research Evaluation*, 14: 186–98.

Gläser, J., and Laudel, G. (2007). ′The Social Construction of Bibliometric Evaluations′, In: Whitley, R. and Gläser, J. (eds) *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, Vol. 26. The Netherlands: Springer, pp. 101–23.

Graber, M., Launov, A., and Wälde, K. (2008) 'Publish or Perish? The Increasing Importance of Publications for Prospective Economics Professors in Austria, Germany and Switzerland', *German Economic Review*, 9/4: 457–72.

Gross, P. L. K., and Gross, E. M. (1927) 'College Libraries and Chemical Education', *Science*, 66: 385–9.

Hammarfelt, B., de Rijcke, S., and Rushforth, A. (2016) 'Quantified Academic Selves: The Gamification of Research through Social Networking Services', *Information Research*, 21/2.

Hargens, L. L., and Schuman, H. (1990) 'Citation Counts and Social Comparisons: Scientists' use and Evaluation of Citation Index Data', *Social Science Research*, 19/3: 205–21.

Haucap, J., and Muck, J. (2015) 'What Drives the Relevance and Reputation of Economics Journals? An Update from a Survey among Economists', *Scientometrics*, 103/3: 849–77.

Hicks, D. (2004) 'The Four Literatures of Social Science', In: Moed, H., Glänzel, W. and Schmoch, U. (eds) *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, pp. 473–96. Dordrecht: Springer.

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: 16569–72.

Karpik, L. (2010). *Valuing the Unique: The Economics of Singularities*. Princeton, NJ: Princeton University Press.

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Larivière, V., Lozano, G. A., and Gingras, Y. (2014) 'Are Elite Journals Declining?', *Journal of the Association for Information Science and Technology*, 65/4: 649–55.

Leydesdorff, L., Wouters, F., and Bornman, L. (2016). 'Professional and Citizen Bibliometrics: Complementarities and Ambivalences in the Development and use of Indicators', *Scientometrics*, 109: p. 2129–2150.

Macfarlane, B. (2007) 'Defining and Rewarding Academic Citizenship: The Implications for University Promotions', *Journal of Higher Education Policy and Management*, 29/3: 261–73.

Maeße, J. (2017), 'The Elitism Dispositif: Hierarchization, Discourses of Excellence and Organizational Change in European Economics', *Higher Education*, 73: p. 909–927.

Malsch, B., and Tessier, S. (2015) 'Journal Ranking Effects on Junior Academics: Identity Fragmentation and Politicization', *Critical Perspectives on Accounting*, 26: 84–98.

Moed, H. F. (2007) 'The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review', *Science and Public Policy*, 34/8: 575–83.

Musselin, C. (2009). *The Market for Academics*. London: Routledge.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, DC: Computer Horizons.

Nilsson, R. (2009). *God vetenskap. Hur forskares vetenskapsuppfattningar uttryckta i sakkunnigutlåtanden förändras i tre skilda discipliner*. Göteborg: Acta Universitatis Gothoburgensis.

Porter, T. M. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Rushforth, A., and de Rijcke, S. (2015) 'Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands', *Minerva*, 53/2: 117–39.

Tourish, D., and Willmott, H. (2015) 'In Defiance of Folly: Journal Rankings, Mindless Measures and the ABS Guide', *Critical Perspectives on Accounting*, 26: 37–46.

Van Eck, N. J. et al. (2013) 'Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research', *PloS One*, 8/4: e62395.

Van Leeuwen, T. (2008) 'Testing the Validity of the Hirsch-index for Research Assessment Purposes', *Research Evaluation*, 17/2: 157–60.

van Raan, A. (1996) 'Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises', *Scientometrics*, 36/3: 397–420.

Waltman, L., and Van Eck, N. J. (2012) 'The Inconsistency of the h-index', *Journal of the American Society for Information Science and Technology*, 63/2: 406–15.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University.

Whitley, R., and Gläser, J. (2008). *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, Vol. 26. Dordrecht: Springer.

Wildgaard, L., Schneider, J. W., and Larsen, B. (2014) 'A Review of the Characteristics of 108 Author-Level Bibliometric Indicators', *Scientometrics*, 101/1: 125–58.

Woolgar, S. (1991) 'Beyond the Citation Debate: Towards a Sociology of Measurement Technologies and their use in Science Policy', *Science and Public Policy*, 18/5: 319–26.

Wouters, P. et al. (2015) *The Metric Tide: Literature Review*. HEFCE.